BMJ Open

Validation of Chronic Obstructive Pulmonary Disease (COPD) Diagnoses in Administrative Databases: A Systematic Review Protocol

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2016-011777
Article Type:	Protocol
Date Submitted by the Author:	03-Mar-2016
Complete List of Authors:	Rimland, Joseph; Italian National Research Center on Aging (INRCA), Geriatrics and Geriatric Emergency Care Abraha, Iosief; Regional Health Authority of Umbria, Health Planning Service Luchetta, Maria Laura; Azienda USL Umbria 1, General Medicine Cozzolino, Francesco; Regional Health Authority of Umbria, Health Planning Service Orso, Massimiliano; Regional Health Authority of Umbria, Health Planning Service Cherubini, Antonio; Italian National Research Center on Aging (INRCA), Geriatrics and Geriatric Emergency Care Dell'Aquila, Giuseppina; Italian National Research Center on Aging (INRCA), Geriatrics and Geriatric Emergency Care Chiatti, Carlos; Italian National Research Centre on Aging (INRCA), Italy Ambrosio, Giuseppe; University of Perugia School of Medicine, Cardiology; Ospedale S. Maria della Misericordia, Medical Administration Montedori, Alessandro; Regional Health Authority of Umbria, Health Planning Service
Primary Subject Heading:	Research methods
Secondary Subject Heading:	Respiratory medicine, Public health, Epidemiology
Keywords:	Administrative database, ICD-9, ICD-10, Chronic pulmonary obstructive disease, sensitivity and specificity, systematic review

Validation of Chronic Obstructive Pulmonary Disease (COPD) Diagnoses in Administrative Databases: A Systematic Review Protocol

Joseph M Rimland, Iosief Abraha, Maria Laura Luchetta, Francesco Cozzolino, Massimiliano Orso, Antonio Cherubini, Giuseppina Dell'Aquila, Carlos Chiatti, Giuseppe Ambrosio, Alessandro Montedori.

Author affiliations:
Health Planning Service, Regional Health Authority of Umbria, Perugia, Italy
Iosief Abraha
Alessandro Montedori
Francesco Cozzolino
Massimiliano Orso

Scientific Directorate, Italian National Research Center on Aging, Ancona, Italy
Carlos Chiatti

Geriatrics and Geriatric Emergency Care, Italian National Research Center on Aging, Ancona, Italy
Antonio Cherubini
Joseph M Rimland
Giuseppina Dell'Aquila

Azienda USL Umbria 1, General Medicine, Perugia, Italy
Maria Laura Luchetta

University of Perugia School of Medicine, Cardiology, Perugia, Italy
Giuseppe Ambrosio

Correspondence to:
Dr. Iosief Abraha
Health Planning Service
Regional Health Authority of Umbria
Via Mario Angeloni, 61
06124 Perugia (Italy)
tel +39 075 504 5251
cell. +39349 077 0910
fax +39 075 504 5569
e-mail: iosief_a@yahoo.it
iabraha@regione.umbria.it

Abstract

Introduction Administrative healthcare databases are useful sources to investigate the epidemiology of chronic obstructive pulmonary disease (COPD), to assess longitudinal outcomes in subjects with COPD, and to develop disease management strategies. However, in order to constitute a reliable source for research, administrative databases need to be validated. The aim of this protocol is to perform the first systematic review of studies reporting the validation of *International Classification of Diseases 9th Revision* and *10th Revision* (ICD-9; ICD-10) codes for COPD diagnoses in administrative healthcare databases.

Methods and analysis MEDLINE, EMBASE, Web of Science and the Cochrane Library databases will be searched, using appropriate search strategies. Studies that evaluated the validity of COPD codes in administrative data or studies that used administrative databases to identify COPD diagnoses will be included. Inclusion criteria will be: (a) the presence of a reference standard case definition for COPD; (b) the presence of at least one test measure (e.g. sensitivity, positive predictive values, etc.); and (c) the use of an administrative database as a data source. Pairs of reviewers will independently abstract data using standardised forms and will assess quality using a checklist based on the Standards for Reporting of Diagnostic accuracy (STARD) criteria. This systematic review protocol has been produced in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses Protocol (PRISMA-P) 2015 statement.

Ethics and dissemination Ethics approval is not required. Results of this study will be submitted to a peer-reviewed journal for publication. The results from this systematic review will be used for outcome research on COPD and will serve as a guide to identify appropriate case definitions of COPD, and reference standards, for researchers involved in validating administrative healthcare databases.

Trial registration number PROSPERO 2015 CRD42015029204

Strengths and limitations of this study

- Validation of *International Classification of Diseases 9th Revision* and *10th Revision* (ICD-9; ICD-10) diagnosis codes for Chronic Obstructive Pulmonary Disease using administrative healthcare databases can contribute to health outcome research.
- This review will be the first to systematically identify and evaluate primary studies that validated the accuracy of administrative healthcare databases with ICD-9 and ICD-10 codes for Chronic Obstructive Pulmonary Disease.

Introduction

Chronic Obstructive Pulmonary Disease (COPD) is a diverse collection of lung diseases including chronic bronchitis, emphysema and chronic obstructive airway disease. It is distinguished by continuous airflow restriction, is frequently progressive and is associated with a chronically increased airway and lung inflammatory reaction to gases or particles [1 2]. COPD is correlated with significant morbidity and mortality and is the fourth leading cause of death worldwide [3]. On the basis of WHO estimates (2004), 64 million people had moderate to severe COPD, which led to 3 million deaths [4]. The burden of COPD is estimated to increase in the near future, because of continued exposure to risk factors and ageing of the population [1 2]. Smoking is the main cause of COPD, but other factors, especially exposure to occupational or environmental airborne irritants, may also contribute to the development of this group of lung diseases [1 2].

Administrative healthcare databases are increasingly being used to examine features of health care delivery, including practice patterns, quality of care, safety and efficacy of drugs, and epidemiological studies. Some of the advantages included the minimisation of recall bias, better generalizability than randomised trials and better cost-effectiveness approach to research compared to primary data collection[5]. To be reliably used for research, administrative healthcare databases need to be validated concerning the disease of interest[6-9]. This means that the content of the databases (e.g., a codes of a disease) need to be ascertained using a reference standard (e.g., medical chart)[10]. Alternatively, algorithms can be developed by combining multiple codes – or sets of codes (e.g., diagnosis codes plus prescription or spirometry data) to enhance the ability to identify events of interest in the database [10-14]. Administrative healthcare databases are excellent resources to determine the epidemiology [14-16], and burden of COPD [17 18] and to evaluate longitudinal outcomes of the disease [19 20]. Results from analysing these healthcare databases can assist in developing disease management strategies (including education regarding the disease, optimisation of evidence-based medications, information, case manager support and institution of self-management principles) to improve the health of subjects suffering from COPD [21].

The current *International Classification of Diseases, 9th Revision* (ICD-9) codes for COPD are 491, 492 or 496, while the corresponding *International Classification of Diseases, 10th Revision* (ICD-10) codes are J42, J43 and J44.

There are several studies that assessed the validity of administrative databases for COPD [10 14 22], however, to our knowledge, no systematic assessment of algorithms or case definitions of COPD have been performed in the medical literature. With the present protocol, we aim to systematically evaluate validation studies of diagnostic codes or algorithms to identify cases of COPD.

Research question

The primary research question is the accuracy of algorithms to correctly identify patients with COPD in administrative databases. The target populations are patients with COPD, the index test will be administrative data algorithms for COPD, the reference standard will be medical charts, validated electronic health records or COPD registries. Our primary outcome is the accuracy (expressed in terms of sensitivity, specificity and positive and negative predictive values) of administrative data algorithms to discriminate cases of COPD.

Methods

Literature search

Comprehensive searches of MEDLINE, EMBASE, the Web of Science and the Cochrane Library from their inception, will be performed to identify published peer-reviewed articles. A search strategy will be employed that we developed based on the combination of: (a) keywords and MeSH terms to identify records concerning COPD; and (b) a search strategy, based on the combination of terms used by Benchimol et al. [23] and the Mini-Sentinel's program [24 25], which is designed to accurately capture studies that use healthcare administrative databases. The developed search strategy is available as supplementary material (**Appendix**). To retrieve additional articles, relevant reference lists of key articles will be hand searched. The "Cited-By" tools in PubMed and Google Scholar will also be used to find relevant articles that cited the article of interest, identified through the aforementioned search strategy. Titles and abstracts will be screened for eligibility by two independent reviewers and discrepancies will be resolved by discussion.

This review protocol has been prepared according to the Preferred Reporting Items for Systematic reviews and Meta-Analysis Protocols (PRISMA-P) 2015 Statement [26] and the results will be presented following the PRISMA flow diagram (**Figure**). This protocol has also been published in the PROSPERO International Prospective Register of Systematic Reviews with registration number CRD42015029204 (http://www.crd.york.ac.uk/PROSPERO).

Inclusion criteria

Full-texts of eligible peer-reviewed articles, without limits on publication date and published in English, that used administrative data for the ICD-9 or ICD-10 codes for COPD diagnoses, will be obtained. For each study, the following inclusion criteria will be applied: (a) the presence of a reference standard case definition for the disease of interest; (b) the presence of at least one test measure (e.g., sensitivity, positive predictive values, etc.); (c) the use of an administrative database

(i.e., a database in which data is routinely and passively collected without an a priori research question) as a data source; and (d) the use of a database from a representative sample of the general population. Studies that used electronic health records (EHR) to validate COPD will also be included. [12] [27]. Conversely, studies that employed databases that were not truly administrative (e.g. disease registries, epidemiology surveillance systems, etc.) will be excluded.

At the initial stage, titles and abstracts will be screened for potentially eligible studies. Subsequently, full texts of articles will be obtained and assessed to determine if they meet the inclusion and exclusion criteria. Data abstraction will be conducted using standardised data collection forms, which will first be tested on a sample of eligible articles. Two review authors working independently, and in duplicate, will carry out title, abstract and full-text screening and data abstraction. Any discrepancies will be resolved by consensus, and where necessary, a third review author will be involved. Calibration exercises will be performed at each level of the process.

Data extraction

Data extraction will include the following information:

(a) the details of the included study (containing the title, the year of publication and the journal, the country of origin, and the sources of funding; the first author will be used as the study ID);

(b) the disease of interest (COPD);

(c) the target population from which the administrative data were collected;

(d) the type of administrative database used (e.g., hospitalisation discharge data), outpatient records (e.g., physician billing claims, etc.);

(e) the ICD-9 or ICD-10 code used or the administrative data algorithm(s) tested (including prescriptions fills, Current Procedural Terminology, timing of diagnosis, etc.)

(f) the modality of algorithm development (e.g., using Classification and Regression Trees, logistic regression, expert opinion…);

(g) external validation;

(h) the use of training and testing cohorts;

(i) the reference standard used to determine the validity of the diagnostic code (e.g., medical chart review, patient self-reports, disease registry, etc.);

(j) the characteristic of the test used to determine the validity of the diagnostic code or algorithm (e.g., sensitivity, specificity, positive predictive values (PPVs) and negative predictive values (NPVs), area under the receiver operating characteristic curve, likelihood ratios, and kappa statistics);

(k) any funding source and conflict of interest.

Quality assessment

The design and methods of the included primary studies will be assessed using a checklist developed by Benchimol et al. [23], based on the criteria published by the Standards for Reporting of Diagnostic accuracy (STARD) initiative for the accurate reporting of diagnostic studies [28]. The checklist is based on a standardized 40-item which it is possible to assess the quality of the methods and reporting of studies that validated codes or algorithms used to identify patients with the disease of interest within an administrative database (**Appendix**). Two reviewers will be involved in the quality assessment and will work in duplicate and independently. Any disagreement will be solved by discussion. The presence of potential biases within the studies will be reported descriptively.

No subgroup analysis or publication bias assessment are anticipated.

Analysis

For each algorithm, the performance statistics, provided in each of the included studies, will be abstracted. Validation statistics may include sensitivity, specificity, PPV and NPV. Sensitivity measures the degree to which an ICD-9 or ICD-10 code (e.g., ICD-9 491) correctly identifies individuals possessing the characteristic of interest (i.e., COPD) in the source used as a reference

standard (e.g. medical chart) [29]. PPV is the number of true positives divided by the total number of cases receiving the code and expresses the likelihood that the code corresponds to a true-positive case. NPV is the number of true negatives divided by the total number of cases without the code of interest and expresses the likelihood that the absence of the code corresponds to a true-negative case. Where possible, PPVs and NPVs will be calculated if not reported. Ninety-five percent confidence intervals (95% CI) will be calculated when they are not reported in the articles. Where possible, validation statistics will be aggregated and stratified by administrative data source (outpatient vs. inpatient data), type of ICD code (ICD-9 or ICD-10), and country of origin.

Ethics and dissemination

This review protocol will use publicly available data without directly involving human participants, hence approval from an ethics committee is not required. An outline of the protocol has been published in the PROSPERO International Prospective Register of Systematic Reviews in 2015, registration number CRD42015029204. The results will summarise the studies that validated diagnostic codes for Chronic Obstructive Pulmonary Disease in administrative data. Where possible, a quantitative synthesis of the accuracy data will be provided and the outcomes using different algorithms will be discussed. Findings of the review will be presented at relevant scientific conferences and disseminated through publication in a peer-reviewed journal.

Footnotes

Contributors IA, JMR, and MLL conceived the study. JMR, IA, MLL, FC, MO, AC, GD, CC, GA, and AM were responsible for designing the protocol. MLL, JMR and IA drafted the protocol manuscript. JMR, IA, FC, and MO developed the search strategy. JMR, IA, MLL, FC, MO, AC, GD, CC, GA, and AM critically revised the successive versions of the manuscript and approved the final version. IA acts as guarantor.

Funding Currently the study is not funded.

Competing interests None.

References

1. Jones PW, Nadeau G, Small M, Adamek L. Characteristics of a COPD population categorised using the GOLD framework by health status and exacerbations. Respiratory Medicine 2014;**108**(1):129-35 doi: http://dx.doi.org/10.1016/j.rmed.2013.08.015[published Online First: Epub Date]|.
2. Rabe KF, Hurd S, Anzueto A, et al. Global Strategy for the Diagnosis, Management, and Prevention of Chronic Obstructive Pulmonary Disease. American Journal of Respiratory and Critical Care Medicine 2007;**176**(6):532-55 doi: 10.1164/rccm.200703-456SO[published Online First: Epub Date]|.
3. Lange P, Marott JL, Vestbo J, et al. Prediction of the Clinical Course of Chronic Obstructive Pulmonary Disease, Using the New GOLD Classification. American Journal of Respiratory and Critical Care Medicine 2012;**186**(10):975-81 doi: 10.1164/rccm.201207-1299OC[published Online First: Epub Date]|.
4. Adeloye D, Chua S, Lee C, et al. Global and regional estimates of COPD prevalence: Systematic review and meta-analysis. Journal of global health 2015;**5**(2):020415 doi: 10.7189/jogh.05-020415[published Online First: Epub Date]|.
5. Schneeweiss S, Avorn J. A review of uses of health care utilization databases for epidemiologic research on therapeutics. Journal of clinical epidemiology 2005;**58**(4):323-37 doi: 10.1016/j.jclinepi.2004.10.012[published Online First: Epub Date]|.
6. Abraha I, Giovannini G, Serraino D, Fusco M, Montedori A. Validity of Breast, Lung and Colorectal Cancer Diagnoses in Administrative Databases: A Systematic Review Protocol. BMJ Open 2016:010409
7. Abraha I, Montedori A, Eusebi P, et al. The Current State of Validation of Administrative Healthcare Databases in Italy: A Systematic Review. Pharmacoepidemiology and Drug Safety 2012;**21**:400-00
8. Abraha I, Serraino D, Giovannini G, et al. Validity of ICD-9-CM Codes for Breast, Lung, and Colorectal Cancers in Three Italian Administrative Healthcare Databases: A Diagnostic Accuracy Study Protocol BMJ Open 2016;**Accept (22-Feb-2016)**:010547
9. West SL, Ritchey ME, Poole C. Validity of Pharmacoepidemiologic Drug and Diagnosis Data. Pharmacoepidemiology: Wiley-Blackwell, 2012:757-94.
10. Lacasse Y, Daigle JM, Martin S, Maltais F. Validity of chronic obstructive pulmonary disease diagnoses in a large administrative database. Canadian respiratory journal : journal of the Canadian Thoracic Society 2012;**19**(2):e5-9
11. Lix LM, De Coster C, Currie R. *Defining and validating chronic diseases: an administrative data approach*: Manitoba Centre for Health Policy Winnipeg, 2006.
12. Chubak J, Pocobelli G, Weiss NS. Tradeoffs between accuracy measures for electronic health care data algorithms. Journal of clinical epidemiology 2012;**65**(3):343-49.e2 doi: 10.1016/j.jclinepi.2011.09.002[published Online First: Epub Date]|.
13. Sharifi M, Krishanswami S, McPheeters ML. A systematic review of validated methods to capture acute bronchospasm using administrative or claims data. Vaccine 2013;**31 Suppl 10**:K12-20 doi: 10.1016/j.vaccine.2013.06.091[published Online First: Epub Date]|.
14. Green ME, Natajaran N, O'Donnell DE, et al. Chronic obstructive pulmonary disease in primary care: an epidemiologic cohort study from the Canadian Primary Care Sentinel Surveillance Network. CMAJ open 2015;**3**(1):E15-E22 doi: 10.9778/cmajo.20140040[published Online First: Epub Date]|.
15. Anecchino C, Rossi E, Fanizza C, De Rosa M, Tognoni G, Romero M. Prevalence of chronic obstructive pulmonary disease and pattern of comorbidities in a general population. International Journal of Chronic Obstructive Pulmonary Disease 2007;**2**(4):567-74
16. Faustini A, Canova C, Cascini S, et al. The reliability of hospital and pharmaceutical data to assess prevalent cases of chronic obstructive pulmonary disease. Copd 2012;**9**(2):184-96 doi: 10.3109/15412555.2011.654014[published Online First: Epub Date]|.
17. Bustacchini S, Chiatti C, Furneri G, Lattanzio F, Mantovani LG. The economic burden of chronic obstructive pulmonary disease in the elderly: results from a systematic review of the literature. Current opinion in pulmonary medicine 2011;**17 Suppl 1**:S35-41 doi: 10.1097/01.mcp.0000410746.82840.79[published Online First: Epub Date]|.

18. Simoni-Wastila L, Blanchette CM, Qian J, et al. Burden of chronic obstructive pulmonary disease in Medicare beneficiaries residing in long-term care facilities. The American journal of geriatric pharmacotherapy 2009;**7**(5):262-70 doi: 10.1016/j.amjopharm.2009.11.003[published Online First: Epub Date]|.
19. Ismaila A, Corriveau D, Vaillancourt J, et al. Impact of adherence to treatment with tiotropium and fluticasone propionate/salmeterol in chronic obstructive pulmonary diseases patients. Current medical research and opinion 2014;**30**(7):1427-36 doi: 10.1185/03007995.2014.908828[published Online First: Epub Date]|.
20. Abudagga A, Sun SX, Tan H, Solem CT. Exacerbations among chronic bronchitis patients treated with maintenance medications from a US managed care population: an administrative claims data analysis. Int J Chron Obstruct Pulmon Dis 2013;**8**:175-85 doi: 10.2147/copd.s40437[published Online First: Epub Date]|.
21. Rice KL, Dewan N, Bloomfield HE, et al. Disease management program for chronic obstructive pulmonary disease: a randomized controlled trial. American journal of respiratory and critical care medicine 2010;**182**(7):890-6 doi: 10.1164/rccm.200910-1579OC[published Online First: Epub Date]|.
22. Gershon AS, Wang C, Guan J, Vasilevska-Ristovska J, Cicutto L, To T. Identifying individuals with physcian diagnosed COPD in health administrative databases. Copd 2009;**6**(5):388-94
23. Benchimol EI, Manuel DG, To T, Griffiths AM, Rabeneck L, Guttmann A. Development and use of reporting guidelines for assessing the quality of validation studies of health administrative data. Journal of clinical epidemiology 2011;**64**(8):821-9 doi: 10.1016/j.jclinepi.2010.10.006[published Online First: Epub Date]|.
24. Carnahan RM, Moores KG. Mini-Sentinel's systematic reviews of validated methods for identifying health outcomes using administrative and claims data: methods and lessons learned. Pharmacoepidemiol Drug Saf 2012;**21 Suppl 1**:82-9 doi: 10.1002/pds.2321[published Online First: Epub Date]|.
25. McPheeters ML, Sathe NA, Jerome RN, Carnahan RM. Methods for systematic reviews of administrative database studies capturing health outcomes of interest. Vaccine 2013;**31 Suppl 10**:K2-6 doi: 10.1016/j.vaccine.2013.06.048[published Online First: Epub Date]|.
26. Shamseer L, Moher D, Clarke M, et al. Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015: elaboration and explanation. Bmj 2015;**349**:g7647 doi: 10.1136/bmj.g7647[published Online First: Epub Date]|.
27. Dean BB, Lam J, Natoli JL, Butler Q, Aguilar D, Nordyke RJ. Review: Use of Electronic Medical Records for Health Outcomes Research: A Literature Review. Medical Care Research and Review 2009;**66**(6):611-38 doi: 10.1177/1077558709332440[published Online First: Epub Date]|.
28. Bossuyt PM, Reitsma JB, Bruns DE, et al. Towards complete and accurate reporting of studies of diagnostic accuracy: The STARD Initiative. Annals of internal medicine 2003;**138**(1):40-4
29. West SL, Strom BL, Poole C. *Validity of Pharmacoepidemiologic Drug and Diagnosis Data*: John Wiley & Sons, Ltd, 2007.

Figure . Study screening process (PRISMA Flow Diagram)

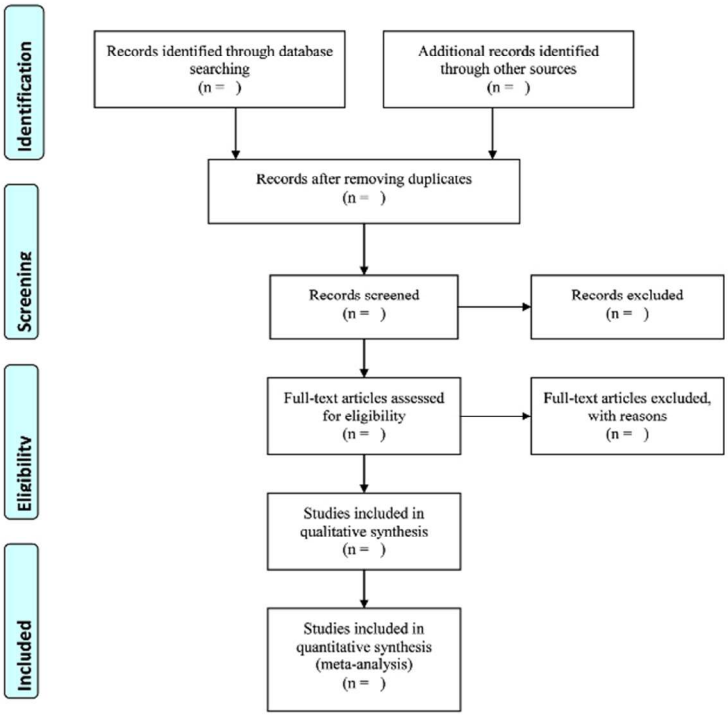

Figure 1. Study screening process
61x86mm (300 x 300 DPI)

Appendix 1

MEDLINE (via Pubmed) search strategy

1. (health administrative) OR (administrative data) OR (administrative database) OR (claim administrative) OR (International Classification of Diseases) OR "International Classification of Diseases"[Mesh] OR ICD-9-CM OR ICD-10 OR "Database Management Systems"[Mesh] OR "Medical Records Systems, Computerized"[Mesh] OR "CPT" OR "Current procedural terminology"[Mesh]
2. (factual databases) OR (geographic information systems) OR (national practitioner data bank) OR (insurance database)
3. #1 OR #2
4. sensitivity or "Sensitivity and Specificity"[Mesh]
5. specificity[Title/Abstract]
6. (positive predictive value) OR (negative predictive value) OR (likelihood ratio) OR (receiver operating characteristic) OR kappa
7. ((case or cases) AND (verificat* OR valid* OR identif* OR definition* OR define* OR evaluat*))
8. Algorithm OR "Algorithm"[Mesh]
9. #4 OR #5 OR #6 OR #7 OR #8
10. emphysema
11. (chronic* bronchitis*)
12. (obstruct* (pulmonary or lung* or airway* or airflow* or bronch* or respirat*))
13. (chronic obstructive pulmonary disease)[Text Word]
14. COPD
15. CAPD
16. "pulmonary disease, chronic obstructive"[MeSH Terms]
17. #10 OR #11 OR #12 OR #13 OR #14 OR #15 OR #16
18. #3 AND #9 AND #17

EMBASE search strategy (via embase.com)

1. health NEAR/3 administrative OR administrative NEAR/3 data OR administrative NEAR/3 database OR claim NEAR/3 administrative OR (International Classification of Diseases) OR 'International Classification of Diseases'/exp OR ICD-9-CM OR ICD-10 OR 'Database Management Systems'/exp OR 'Medical Records Systems, Computerized'/exp OR 'CPT' OR 'Current procedural terminology'/exp
2. database:ab,ti OR (('practitioner'/exp OR practitioner) AND data AND bank) OR (('practitioner'/exp OR practitioner) AND ('database'/exp OR database)) OR ('insurance' AND ('database'/exp OR database))
3. #1 OR #2
4. 'sensitivity and specificity'/exp OR 'sensitivity and specificity'
5. specificity:ab,ti
6. 'predictive value of tests'/exp OR 'predictive value of tests'
7. (positive:ab,ti AND predictive:ab,ti AND value:ab,ti) OR (negative:ab,ti AND predictive:ab,ti AND value:ab,ti) OR (likelyhood:ab,ti AND ratio:ab,ti) OR (receiver:ab,ti AND operating:ab,ti AND characteristic:ab,ti) OR kappa:ab,ti
8. case NEAR/1 (verificat* OR valid* OR identif* OR definition* OR define* OR evaluat*)
9. 'algorithms'/exp OR algorithm
10. #4 OR #5 OR #6 OR #7 OR #8 OR #9
11. 'emphysema'/exp
12. 'chronic bronchitis'/exp
13. (obstruct* NEAR/3 (pulmonary or lung* or airway* or airflow* or bronch* or respirat*))
14. 'chronic obstructive pulmonary disease '/exp
15. 'COPD'/exp
16. 'CAPD'/exp
17. #11 OR #12 OR #13 OR #14 OR #15 OR #16
18. #3 AND #10 AND #17

Web of Science search strategy

1. (health NEAR/3 administrative) OR (administrative NEAR/3 data) OR (administrative NEAR/3 database) OR (claim NEAR/3 administrative) OR (International Classification of Diseases) OR ICD-9-CM OR ICD-10 OR (Database Management Systems) OR ("Medical Records Systems" NEAR/2 Computerized) OR "CPT" OR (Current procedural terminology)
2. (factual databases) OR (geographic information systems) OR (national practitioner data bank) OR (insurance database)
3. #1 OR #2
4. sensitivity or "Sensitivity and Specificity"
5. specificity
6. (positive predictive value) OR (negative predictive value) OR (likelihood ratio) OR (receiver operating characteristic) OR kappa
7. ((case or cases) AND (verificat* OR valid* OR identif* OR definition* OR define* OR evaluat*))
8. algorithm
9. #4 OR #5 OR #6 OR #7 OR #8
10. emphysema
11. (chronic* NEAR/3 bronchitis*)
12. (obstruct* NEAR/3 (pulmonary or lung* or airway* or airflow* or bronch* or respirat*))
13. chronic obstructive pulmonary disease
14. (COPD)
15. (CAPD)
16. #10 OR #11 OR #12 OR #13 OR #14 OR #15
17. # 3 AND #9 AND #16

The Cochrane Library

1. (health near/3 administrative) or (administrative near/3 data) or (administrative near/3 database) or (claim near/3 administrative) or (International Classification of Diseases) or [mh "International Classification of Diseases"] or ICD-9-CM or ICD-10 or [mh "Database Management Systems"] or [mh "Medical Records Systems, Computerized"] or "CPT" or [mh "Current procedural terminology"]
2. (factual databases) OR (geographic information systems) OR (national practitioner data bank) OR (insurance database)
3. #1 OR #2
4. sensitivity or [mh "Sensitivity and Specificity"]
5. specificity:ti,ab,kw
6. (positive predictive value) OR (negative predictive value) OR (likelihood ratio) OR (receiver operating characteristic) OR kappa
7. ((case or cases) AND (verificat* OR valid* OR identif* OR definition* OR define* OR evaluat*))
8. Algorithm OR [mh "Algorithm"]
9. #4 OR #5 OR#6 OR #7 OR #8
10. emphysema
11. (chronic* near/3 bronchitis*)
12. (obstruct* near/3 (pulmonary or lung* or airway* or airflow* or bronch* or respirat*))
13. (chronic obstructive pulmonary disease)
14. COPD
15. CAPD
16. [mh "pulmonary disease, chronic obstructive"]
17. #10 OR #11 OR #12 OR #13 OR #14 OR #15 OR #16
18. # 3 AND #9 AND #17

Appendix 2

Checklist of reporting criteria for studies validating health administrative data algorithms (developed by Benchimol et al., based on the criteria published by the Standards for Reporting of Diagnostic accuracy (STARD) initiative for the accurate reporting of studies using diagnostic studies.

	YES	NO	UNCERTAIN	NOT APPLICABLE
TITLE, KEYWORDS, ABSTRACT				
Identify article as study of assessing diagnostic accuracy				
Identify article as study of administrative data				
INTRODUCTION:				
State disease identification & validation one of goals of study				
METHODS:				
Participants in validation cohort:				
Describe validation cohort (Cohort of patients to which reference standard was applied)				
• Age				
• Disease				
• Severity				
• Location/Jurisdiction				
Describe recruitment procedure of validation cohort				
• Inclusion criteria				
• Exclusion criteria				
Describe patient sampling (random, consecutive, all, etc.)				
Describe data collection				
• Who identified patients and did selection adhere to patient recruitment criteria				
• Who collected data				
• *A priori* data collection form				
• Disease classification				
• Split sample (i.e. re-validation using a separate cohort) a) Training set b) Testing set				
Test Methods:				
Describe number, training and expertise of persons reading reference standard				
If >1 person reading reference standard, quote measure of consistency (e.g. kappa)				

Blinding of interpreters of reference standard to results of classification by administrative data e.g. Chart abstractor blinded to how that chart was coded				
Statistical Methods:				
Describe methods of calculating/comparing diagnostic accuracy				
RESULTS:				
Participants:				
Report when study done, start/end dates of enrollment				
Describe number of people who satisfied inclusion/exclusion criteria				
Study flow diagram				
Test results:				
Report distribution of disease severity				
Report cross-tabulation of index tests by results of reference standard				
Estimates:				
Report at least 4 estimates of diagnostic accuracy				
Diagnostic Accuracy Measures Reported:				
• Sensitivity				
• Spec				
• PPV				
• NPV				
• Likelihood ratios				
• Kappa				
• Area under the ROC curve / c-statistic				
• Accuracy/agreement				
• Other (specify)				
Report accuracy for subgroups (e.g. age, geography, different sex, etc.)				
If PPV/NPV reported, ratio of cases/controls of validation cohort approximate prevalence of condition in the population				
Report 95% confidence intervals for each diagnostic measure				
DISCUSSION:				
Discuss the applicability of the validation findings				

PRISMA-P (Preferred Reporting Items for Systematic review and Meta-Analysis Protocols) 2015 checklist: recommended items to address in a systematic review protocol*

Section and topic	Item No	Checklist item	Page
ADMINISTRATIVE INFORMATION			
Title:			
Identification	1a	Identify the report as a protocol of a systematic review	Page 1
Update	1b	If the protocol is for an update of a previous systematic review, identify as such	
Registration	2	If registered, provide the name of the registry (such as PROSPERO) and registration number	Page 2: Trial registration number PROSPERO 2015 CRD42015029204
Authors:			
Contact	3a	Provide name, institutional affiliation, e-mail address of all protocol authors; provide physical mailing address of corresponding author	Page 1
Contributions	3b	Describe contributions of protocol authors and identify the guarantor of the review	Page 10
Amendments	4	If the protocol represents an amendment of a previously completed or published protocol, identify as such and list changes; otherwise, state plan for documenting important protocol amendments	At this stage there are no relevant amendments to perform
Support:			
Sources	5a	Indicate sources of financial or other support for the review	Page 10 *not funded*
Sponsor	5b	Provide name for the review funder and/or sponsor	Page 10
Role of sponsor or funder	5c	Describe roles of funder(s), sponsor(s), and/or institution(s), if any, in developing the protocol	Page 10
INTRODUCTION			
Rationale	6	Describe the rationale for the review in the context of what is already known	Page 4 and 5
Objectives	7	Provide an explicit statement of the question(s) the review will address with reference to participants, interventions, comparators, and outcomes (PICO)	Page 5
METHODS			
Eligibility criteria	8	Specify the study characteristics (such as PICO, study design, setting,	Page 5, 6:

		time frame) and report characteristics (such as years considered, language, publication status) to be used as criteria for eligibility for the review	
Information sources	9	Describe all intended information sources (such as electronic databases, contact with study authors, trial registers or other grey literature sources) with planned dates of coverage	Page 6.
Search strategy	10	Present draft of search strategy to be used for at least one electronic database, including planned limits, such that it could be repeated	Appendix 1 in Supplemental file
Study records:			Page 7:
Data management	11a	Describe the mechanism(s) that will be used to manage records and data throughout the review	Page 7:.
Selection process	11b	State the process that will be used for selecting studies (such as two independent reviewers) through each phase of the review (that is, screening, eligibility and inclusion in meta-analysis)	Page 7:
Data collection process	11c	Describe planned method of extracting data from reports (such as piloting forms, done independently, in duplicate), any processes for obtaining and confirming data from investigators	Page 7:
Data items	12	List and define all variables for which data will be sought (such as PICO items, funding sources), any pre-planned data assumptions and simplifications	Page 7/8 .
Outcomes and prioritization	13	List and define all outcomes for which data will be sought, including prioritization of main and additional outcomes, with rationale	Page 5
Risk of bias in individual studies	14	Describe anticipated methods for assessing risk of bias of individual studies, including whether this will be done at the outcome or study level, or both; state how this information will be used in data synthesis	Not applicable. The present review will apply the STARD criteria (Appendix 2 in Supplemental file).
Data synthesis	15a	Describe criteria under which study data will be quantitatively synthesised	No cumulative evidence will be presented.
	15b	If data are appropriate for quantitative synthesis, describe planned summary measures, methods of handling data and methods of combining data from studies, including any planned exploration of consistency (such as I^2, Kendall's τ)	
	15c	Describe any proposed additional analyses (such as sensitivity or subgroup analyses, meta-regression)	
	15d	If quantitative synthesis is not appropriate, describe the type of summary planned	
Meta-bias(es)	16	Specify any planned assessment of meta-bias(es) (such as publication	Not applicable

		bias across studies, selective reporting within studies)	
Confidence in cumulative evidence	17	Describe how the strength of the body of evidence will be assessed (such as GRADE)	The present review will apply the STARD criteria. Page 8 . (Appendix 2 in Supplemental file).

*** It is strongly recommended that this checklist be read in conjunction with the PRISMA-P Explanation and Elaboration (cite when available) for important clarification on the items. Amendments to a review protocol should be tracked and dated. The copyright for PRISMA-P (including checklist) is held by the PRISMA-P Group and is distributed under a Creative Commons Attribution Licence 4.0.**

From: Shamseer L, Moher D, Clarke M, Ghersi D, Liberati A, Petticrew M, Shekelle P, Stewart L, PRISMA-P Group. Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015: elaboration and explanation. BMJ. 2015 Jan 2;349(jan02 1):g7647.

BMJ Open

Validation of Chronic Obstructive Pulmonary Disease (COPD) Diagnoses in Healthcare Databases: A Systematic Review Protocol

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2016-011777.R1
Article Type:	Protocol
Date Submitted by the Author:	30-Apr-2016
Complete List of Authors:	Rimland, Joseph; Italian National Research Center on Aging (INRCA), Geriatrics and Geriatric Emergency Care Abraha, Iosief; Regional Health Authority of Umbria, Health Planning Service Luchetta, Maria Laura; Azienda USL Umbria 1, General Medicine Cozzolino, Francesco; Regional Health Authority of Umbria, Health Planning Service Orso, Massimiliano; Regional Health Authority of Umbria, Health Planning Service Cherubini, Antonio; Italian National Research Center on Aging (INRCA), Geriatrics and Geriatric Emergency Care Dell'Aquila, Giuseppina; Italian National Research Center on Aging (INRCA), Geriatrics and Geriatric Emergency Care Chiatti, Carlos; Italian National Research Centre on Aging (INRCA), Italy Ambrosio, Giuseppe; University of Perugia School of Medicine, Cardiology; Ospedale S. Maria della Misericordia, Medical Administration Montedori, Alessandro; Regional Health Authority of Umbria, Health Planning Service
Primary Subject Heading:	Research methods
Secondary Subject Heading:	Respiratory medicine, Public health, Epidemiology
Keywords:	Administrative database, Chronic pulmonary obstructive disease, sensitivity and specificity, systematic review, Electronic Health Record, COPD algorithm

Validation of Chronic Obstructive Pulmonary Disease (COPD) Diagnoses in Healthcare Databases: A Systematic Review Protocol

Joseph M Rimland, Iosief Abraha, Maria Laura Luchetta, Francesco Cozzolino, Massimiliano Orso, Antonio Cherubini, Giuseppina Dell'Aquila, Carlos Chiatti, Giuseppe Ambrosio, Alessandro Montedori.

Author affiliations:
Health Planning Service, Regional Health Authority of Umbria, Perugia, Italy
Iosief Abraha
Alessandro Montedori
Francesco Cozzolino
Massimiliano Orso

Scientific Directorate, Italian National Research Center on Aging, Ancona, Italy
Carlos Chiatti

Geriatrics and Geriatric Emergency Care, Italian National Research Center on Aging, Ancona, Italy
Antonio Cherubini
Joseph M Rimland
Giuseppina Dell'Aquila

Azienda USL Umbria 1, General Medicine, Perugia, Italy
Maria Laura Luchetta

University of Perugia School of Medicine, Cardiology, Perugia, Italy
Giuseppe Ambrosio

Correspondence to:
Dr. Iosief Abraha
Health Planning Service
Regional Health Authority of Umbria
Via Mario Angeloni, 61
06124 Perugia (Italy)
tel +39 075 504 5251
cell. +39349 077 0910
fax +39 075 504 5569
e-mail: iosief_a@yahoo.it
iabraha@regione.umbria.it

Abstract

Introduction Healthcare databases are useful sources to investigate the epidemiology of chronic obstructive pulmonary disease (COPD), to assess longitudinal outcomes in subjects with COPD, and to develop disease management strategies. However, in order to constitute a reliable source for research, healthcare databases need to be validated. The aim of this protocol is to perform the first systematic review of studies reporting the validation of codes related to COPD diagnoses in healthcare databases.

Methods and analysis MEDLINE, EMBASE, Web of Science and the Cochrane Library databases will be searched, using appropriate search strategies. Studies that evaluated the validity of COPD codes (such as the *International Classification of Diseases 9th Revision* and *10th Revision* system; the *Real codes system* or the *International Classification of Primary Care*) in healthcare databases will be included. Inclusion criteria will be: (a) the presence of a reference standard case definition for COPD; (b) the presence of at least one test measure (e.g. sensitivity, positive predictive values, etc.); and (c) the use of a healthcare database (including administrative claims databases, electronic healthcare databases or COPD registries) as a data source. Pairs of reviewers will independently abstract data using standardised forms and will assess quality using a checklist based on the Standards for Reporting of Diagnostic accuracy (STARD) criteria. This systematic review protocol has been produced in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses Protocol (PRISMA-P) 2015 statement.

Ethics and dissemination Ethics approval is not required. Results of this study will be submitted to a peer-reviewed journal for publication. The results from this systematic review will be used for outcome research on COPD and will serve as a guide to identify appropriate case definitions of COPD, and reference standards, for researchers involved in validating healthcare databases.

Trial registration number PROSPERO 2015 CRD42015029204

Strengths and limitations of this study

- Validation of diagnosis codes for Chronic Obstructive Pulmonary Disease (COPD) using healthcare databases can contribute to health outcome research. The diagnosis codes may include the *International Classification of Diseases 9th Revision* and *10th Revision* (ICD-9; ICD-10) system, the *Real code system* and the *International Classification of Primary Care* system.
- This review will be the first to systematically identify and evaluate primary studies that validated the accuracy of healthcare databases with ICD-9 and ICD-10 codes for COPD
- It is expected that different healthcare databases validate different algorithms to identify COPD. Validated algorithms are context specific and may not be generalizable to other settings.

Introduction

Chronic Obstructive Pulmonary Disease (COPD) is a global health problem [1 2]. It is distinguished by continuous airflow restriction, is frequently progressive and is associated with a chronically increased airway and lung inflammatory reaction to gases or particles [3 4]. COPD is correlated with significant morbidity and mortality and is the fourth leading cause of death worldwide [5]. On the basis of WHO estimates (2004), 64 million people had moderate to severe COPD, which led to 3 million deaths [6]. The burden of COPD is estimated to increase in the near future, because of continued exposure to risk factors and ageing of the population [3 4]. Smoking is the main cause of COPD, but other factors, especially exposure to occupational or environmental airborne irritants, may also contribute to the development of this group of lung diseases [3 4].

Healthcare databases are increasingly being used to examine features of health care delivery, including practice patterns, quality of care, safety and efficacy of drugs, and epidemiological studies. Some of the advantages of healthcare databases included the minimisation of recall bias, better generalizability than randomised trials and better cost-effectiveness approach to research compared to primary data collection[7].To be reliably used for research, healthcare databases need to be validated concerning the disease of interest[8-11]. This means that the content of the databases (e.g., a code of a disease) need to be ascertained using a reference standard (e.g., medical chart)[12]. Alternatively, algorithms can be developed by combining multiple codes – or sets of codes (e.g., diagnosis codes plus prescription or spirometry data) to enhance the ability to identify events of interest in the database [12-16].

Healthcare databases generally encompass administrative claims data and Electronic Health Records (EHR). Administrative claims databases routinely collect data passively, for administrative purposes, for health services delivered by healthcare providers and facilities [17]. The patient information collected includes demographics (name, address, birthdate, gender, and marital status), the dates of healthcare services delivered and charges for the services, diagnostic procedures

performed and healthcare service provider information and in some occasions employment, insurance status, occupational limitations.

Administrative claims databases are excellent resources to investigate the epidemiology [16 18 19], and the burden of COPD [20 21] and to evaluate longitudinal outcomes of a disease [22 23]. Results from analysing these databases can assist in developing disease management strategies (including education regarding the disease, optimisation of evidence-based medications, information, case manager support and institution of self-management principles) to improve the health of subjects suffering from COPD [24].

Electronic Health Records (EHRs) consist of digital files used by healthcare providers for patient care, and unlike administrative claims databases, include clinical notes, medical records, the treatment histories of patients, and prescription records, as well as radiology and laboratory data[25]. Despite EHRs are not established for research purposes, similar to most administrative databases, they are frequently used for healthcare delivery and facilitation of decision-making processes as well as research [25 26].

The Clinical Practice Research Datalink (CPRD), used in the UK, is one such EHR. It is an excellent resource in which to study COPD, as it is based on a large cohort, contains disease severity indicators and long-term follow-up information from a patient's integrated medical history [27-29].

Generally, administrative claims databases use the *International Classification of Diseases, 9th Revision* (ICD-9) codes for COPD (491, 492 or 496), or the *International Classification of Diseases, 10th Revision* (ICD-10) codes (J43 and J44). EHRs such as the UK CPRD database employs the Read code, which is a hierarchical clinical coding system of medical and prescription terms [27]. Some Read codes for COPD are 1001, 9876 and 10863. (See [27] for a list of COPD-related Read codes). The International Classification of Primary Care (ICPC) is another coding system, which is widely used in primary health care and in research[30-32]. The codes for COPD in the ICPC system are R79 and R95.

There are several studies that assessed the validity of healthcare databases for COPD [12 16 27 33], however, to our knowledge, no systematic assessment of algorithms or case definitions of COPD have been published in the medical literature. With the present protocol, we aim to systematically evaluate validation studies of diagnostic codes or algorithms to identify cases of COPD.

Research question

The primary research question is the accuracy of algorithms to correctly identify patients with COPD in healthcare databases (administrative claims, EHR, or COPD registries). The target populations are patients with COPD, the index test will be healthcare data algorithms for COPD, the reference standard will be medical charts, validated electronic health records or COPD registries. Our primary outcome is the accuracy (expressed in terms of sensitivity, specificity and positive and negative predictive values) of healthcare data algorithms to discriminate cases of COPD.

Methods

Literature search

Comprehensive searches of MEDLINE, EMBASE, the Web of Science and the Cochrane Library from their inception, will be performed to identify published peer-reviewed articles. A search strategy will be employed that we developed based on the combination of: (a) keywords and MeSH terms to identify records concerning COPD; and (b) a search strategy, based on the combination of terms used by Benchimol et al. [17], the Mini-Sentinel's program [34 35], and a systematic review that evaluated EHR based primary studies[25]. The developed search strategy is available as supplementary material (**Appendix**). To retrieve additional articles, relevant reference lists of key articles will be hand searched. The "Cited-By" tools in PubMed and Google Scholar will also be used to find relevant articles that cited the article of interest, identified through the aforementioned search strategy. Titles and abstracts will be screened for eligibility by two independent reviewers and discrepancies will be resolved by discussion.

This review protocol has been prepared according to the Preferred Reporting Items for Systematic reviews and Meta-Analysis Protocols (PRISMA-P) 2015 Statement [36] and the results will be presented following the PRISMA flow diagram (**Figure**). This protocol has also been published in the PROSPERO International Prospective Register of Systematic Reviews with registration number CRD42015029204 (http://www.crd.york.ac.uk/PROSPERO).

Inclusion criteria

Full-texts of eligible peer-reviewed articles, without limits on publication date and published in English, that used healthcare data to validate diagnosis codes for COPD diagnoses, will be obtained. For each study, the following inclusion criteria will be applied: (a) the presence of a reference standard case definition for the disease of interest; (b) the presence of at least one test measure (e.g., sensitivity, positive predictive values, etc.); (c) the use of an administrative claims or EHR database

as a data source; and (d) the use of a database from a representative sample of the general population[14] [25].

At the initial stage, titles and abstracts will be screened for potentially eligible studies. Subsequently, full texts of articles will be obtained and assessed to determine if they meet the inclusion and exclusion criteria. Data abstraction will be conducted using standardised data collection forms, which will first be tested on a sample of eligible articles. Two review authors working independently, and in duplicate, will carry out title, abstract and full-text screening and data abstraction. Any discrepancies will be resolved by consensus, and where necessary, a third review author will be involved. Calibration exercises will be performed at each level of the process.

Data extraction

Data extraction will include the following information:

(a) the details of the included study (containing the title, the year of publication and the journal, the country of origin, and the sources of funding; the first author will be used as the study ID);

(b) the disease of interest (COPD);

(c) the code tested (such as ICD-9, ICD-10, or R79 and R95);

(d) the algorithm(s) tested including COPD code, prescription fills (e.g., bronchodilators), use of spirometry, Current Procedural Terminology, timing of diagnosis, etc.;

(e) any information about the performance of the COPD definition/algorithm in sub-populations (e.g. age group, sex, smoking status, GOLD grade of airflow limitation[2], socioeconomic status, WHO Body Mass Index (BMI) category, previous record of asthma diagnosis[27])

(f) the target population from which the healthcare data were collected;

(g) the type of healthcare database used (e.g., hospitalisation discharge data, electronic health record etc.);

(h) the modality of algorithm development (e.g., using Classification and Regression Trees, logistic regression, expert opinion…);

(i) external validation;

(j) the use of training and testing cohorts;

(k) the reference standard used to determine the validity of the diagnostic code (e.g., medical chart review, patient self-reports, disease registry, etc.);

(l) the characteristic of the test used to determine the validity of the diagnostic code or algorithm (e.g., sensitivity, specificity, positive predictive values (PPVs) and negative predictive values (NPVs), area under the receiver operating characteristic curve, likelihood ratios, and kappa statistics);

Quality assessment

The design and methods of the included primary studies will be assessed using a checklist developed by Benchimol et al. [17], based on the criteria published by the Standards for Reporting of Diagnostic accuracy (STARD) initiative for the accurate reporting of diagnostic studies [37]. This standardized checklist is composed of 40 items to assess the quality of the methods and the reporting of studies that validated codes or algorithms used to identify patients with the disease of interest within a healthcare database (**Appendix**). Two reviewers will be involved in the quality assessment and will work in duplicate and independently. Any disagreement will be solved by discussion. The presence of potential biases within the studies will be reported descriptively.

No subgroup analysis or publication bias assessment are anticipated.

Analysis

For each algorithm, the performance statistics, provided in each of the included studies, will be abstracted. Validation statistics may include sensitivity, specificity, PPV and NPV. Sensitivity measures the degree to which a diagnosis code (e.g., ICD-9 491 or Read code 1001) correctly

identifies individuals possessing the characteristic of interest (i.e., COPD) in the source used as a reference standard (e.g. medical chart) [38]. PPV is the number of true positives divided by the total number of cases receiving the code and expresses the likelihood that the code corresponds to a true-positive case. NPV is the number of true negatives divided by the total number of cases without the code of interest and expresses the likelihood that the absence of the code corresponds to a true-negative case. Where possible, PPVs and NPVs will be calculated if not reported. Ninety-five percent confidence intervals (95% CI) will be calculated when they are not reported in the articles. Where possible, validation statistics will be aggregated and stratified by healthcare data source (outpatient vs. inpatient data), type of EHR code (ICD-9, ICD-10, Read, etc.), and country of origin.

Meta-analysis

Where there are studies with homogeneous data, we will use raw data to construct meta-analyses. A bivariate model will be used to derive summary estimates of sensitivity and specificity and their 95% CIs[39]. Data will be analysed using a random-effects model so that sensitivity and specificity are assumed to vary across studies. In addition, summary receiver operating characteristic (ROC) curves will be constructed and pooled estimates of LR+, LR-, and diagnostic odds ratios will be calculated.

Ethics and dissemination

This review protocol will use publicly available data without directly involving human participants, hence approval from an ethics committee is not required. An outline of the protocol has been published in the PROSPERO International Prospective Register of Systematic Reviews in 2015, registration number CRD42015029204. The results will summarise the studies that validated diagnostic codes for Chronic Obstructive Pulmonary Disease in administrative data. Where possible, a quantitative synthesis of the accuracy data will be provided and the outcomes using different algorithms will be discussed. Findings of the review will be presented at relevant scientific conferences and disseminated through publication in a peer-reviewed journal.

Footnotes

Contributors IA, JMR, and MLL conceived the study. JMR, IA, MLL, FC, MO, AC, GD, CC, GA, and AM were responsible for designing the protocol. MLL, JMR and IA drafted the protocol manuscript. JMR, IA, FC, and MO developed the search strategy. JMR, IA, MLL, FC, MO, AC, GD, CC, GA, and AM critically revised the successive versions of the manuscript and approved the final version. IA acts as guarantor.

Funding Currently the study is not funded.

Competing interests None.

References

1. GOLD GIfCLO. Pocket guide to COPD diagnosis, management and prevention. www.goldcopd.com. www.goldcopd.com ed, 2015.
2. Vestbo J, Hurd SS, Agusti AG, et al. Global strategy for the diagnosis, management, and prevention of chronic obstructive pulmonary disease: GOLD executive summary. American journal of respiratory and critical care medicine 2013;**187**(4):347-65 doi: 10.1164/rccm.201204-0596PP[published Online First: Epub Date]|.
3. Jones PW, Nadeau G, Small M, Adamek L. Characteristics of a COPD population categorised using the GOLD framework by health status and exacerbations. Respiratory Medicine 2014;**108**(1):129-35 doi: http://dx.doi.org/10.1016/j.rmed.2013.08.015[published Online First: Epub Date]|.
4. Rabe KF, Hurd S, Anzueto A, et al. Global Strategy for the Diagnosis, Management, and Prevention of Chronic Obstructive Pulmonary Disease. American Journal of Respiratory and Critical Care Medicine 2007;**176**(6):532-55 doi: 10.1164/rccm.200703-456SO[published Online First: Epub Date]|.
5. Lange P, Marott JL, Vestbo J, et al. Prediction of the Clinical Course of Chronic Obstructive Pulmonary Disease, Using the New GOLD Classification. American Journal of Respiratory and Critical Care Medicine 2012;**186**(10):975-81 doi: 10.1164/rccm.201207-1299OC[published Online First: Epub Date]|.
6. Adeloye D, Chua S, Lee C, et al. Global and regional estimates of COPD prevalence: Systematic review and meta-analysis. Journal of global health 2015;**5**(2):020415 doi: 10.7189/jogh.05-020415[published Online First: Epub Date]|.
7. Schneeweiss S, Avorn J. A review of uses of health care utilization databases for epidemiologic research on therapeutics. Journal of clinical epidemiology 2005;**58**(4):323-37 doi: 10.1016/j.jclinepi.2004.10.012[published Online First: Epub Date]|.
8. Abraha I, Giovannini G, Serraino D, Fusco M, Montedori A. Validity of Breast, Lung and Colorectal Cancer Diagnoses in Administrative Databases: A Systematic Review Protocol. BMJ Open 2016:010409
9. Abraha I, Montedori A, Eusebi P, et al. The Current State of Validation of Administrative Healthcare Databases in Italy: A Systematic Review. Pharmacoepidemiology and Drug Safety 2012;**21**:400-00
10. Abraha I, Serraino D, Giovannini G, et al. Validity of ICD-9-CM Codes for Breast, Lung, and Colorectal Cancers in Three Italian Administrative Healthcare Databases: A Diagnostic Accuracy Study Protocol BMJ Open 2016;**Accept (22-Feb-2016)**:010547
11. West SL, Ritchey ME, Poole C. Validity of Pharmacoepidemiologic Drug and Diagnosis Data. Pharmacoepidemiology: Wiley-Blackwell, 2012:757-94.
12. Lacasse Y, Daigle JM, Martin S, Maltais F. Validity of chronic obstructive pulmonary disease diagnoses in a large administrative database. Canadian respiratory journal : journal of the Canadian Thoracic Society 2012;**19**(2):e5-9
13. Lix LM, De Coster C, Currie R. *Defining and validating chronic diseases: an administrative data approach*: Manitoba Centre for Health Policy Winnipeg, 2006.
14. Chubak J, Pocobelli G, Weiss NS. Tradeoffs between accuracy measures for electronic health care data algorithms. Journal of clinical epidemiology 2012;**65**(3):343-49.e2 doi: 10.1016/j.jclinepi.2011.09.002[published Online First: Epub Date]|.
15. Sharifi M, Krishanswami S, McPheeters ML. A systematic review of validated methods to capture acute bronchospasm using administrative or claims data. Vaccine 2013;**31 Suppl 10**:K12-20 doi: 10.1016/j.vaccine.2013.06.091[published Online First: Epub Date]|.
16. Green ME, Natajaran N, O'Donnell DE, et al. Chronic obstructive pulmonary disease in primary care: an epidemiologic cohort study from the Canadian Primary Care Sentinel Surveillance Network. CMAJ open 2015;**3**(1):E15-E22 doi: 10.9778/cmajo.20140040[published Online First: Epub Date]|.
17. Benchimol EI, Manuel DG, To T, Griffiths AM, Rabeneck L, Guttmann A. Development and use of reporting guidelines for assessing the quality of validation studies of health administrative data. Journal of clinical epidemiology 2011;**64**(8):821-9 doi: 10.1016/j.jclinepi.2010.10.006[published Online First: Epub Date]|.

18. Anecchino C, Rossi E, Fanizza C, De Rosa M, Tognoni G, Romero M. Prevalence of chronic obstructive pulmonary disease and pattern of comorbidities in a general population. International Journal of Chronic Obstructive Pulmonary Disease 2007;**2**(4):567-74
19. Faustini A, Canova C, Cascini S, et al. The reliability of hospital and pharmaceutical data to assess prevalent cases of chronic obstructive pulmonary disease. Copd 2012;**9**(2):184-96 doi: 10.3109/15412555.2011.654014[published Online First: Epub Date]|.
20. Bustacchini S, Chiatti C, Furneri G, Lattanzio F, Mantovani LG. The economic burden of chronic obstructive pulmonary disease in the elderly: results from a systematic review of the literature. Current opinion in pulmonary medicine 2011;**17 Suppl 1**:S35-41 doi: 10.1097/01.mcp.0000410746.82840.79[published Online First: Epub Date]|.
21. Simoni-Wastila L, Blanchette CM, Qian J, et al. Burden of chronic obstructive pulmonary disease in Medicare beneficiaries residing in long-term care facilities. The American journal of geriatric pharmacotherapy 2009;**7**(5):262-70 doi: 10.1016/j.amjopharm.2009.11.003[published Online First: Epub Date]|.
22. Ismaila A, Corriveau D, Vaillancourt J, et al. Impact of adherence to treatment with tiotropium and fluticasone propionate/salmeterol in chronic obstructive pulmonary diseases patients. Current medical research and opinion 2014;**30**(7):1427-36 doi: 10.1185/03007995.2014.908828[published Online First: Epub Date]|.
23. Abudagga A, Sun SX, Tan H, Solem CT. Exacerbations among chronic bronchitis patients treated with maintenance medications from a US managed care population: an administrative claims data analysis. Int J Chron Obstruct Pulmon Dis 2013;**8**:175-85 doi: 10.2147/copd.s40437[published Online First: Epub Date]|.
24. Rice KL, Dewan N, Bloomfield HE, et al. Disease management program for chronic obstructive pulmonary disease: a randomized controlled trial. American journal of respiratory and critical care medicine 2010;**182**(7):890-6 doi: 10.1164/rccm.200910-1579OC[published Online First: Epub Date]|.
25. Dean BB, Lam J, Natoli JL, Butler Q, Aguilar D, Nordyke RJ. Review: Use of Electronic Medical Records for Health Outcomes Research: A Literature Review. Medical Care Research and Review 2009;**66**(6):611-38 doi: 10.1177/1077558709332440[published Online First: Epub Date]|.
26. Williams F, Boren SA. The role of the electronic medical record (EMR) in care delivery development in developing countries: a systematic review. Informatics in primary care 2008;**16**(2):139-45
27. Quint JK, Mullerova H, DiSantostefano RL, et al. Validation of chronic obstructive pulmonary disease recording in the Clinical Practice Research Datalink (CPRD-GOLD). BMJ Open 2014;**4**(7):e005540 doi: 10.1136/bmjopen-2014-005540[published Online First: Epub Date]|.
28. Rothnie KJ, Mullerova H, Hurst JR, et al. Validation of the Recording of Acute Exacerbations of COPD in UK Primary Care Electronic Healthcare Records. PloS one 2016;**11**(3):e0151357 doi: 10.1371/journal.pone.0151357[published Online First: Epub Date]|.
29. Wurst KE, St Laurent S, Mullerova H, Davis KJ. Characteristics of patients with COPD newly prescribed a long-acting bronchodilator: a retrospective cohort study. Int J Chron Obstruct Pulmon Dis 2014;**9**:1021-31 doi: 10.2147/copd.s58258[published Online First: Epub Date]|.
30. Soler JK, Okkes I, Oskam S, et al. An international comparative family medicine study of the Transition Project data from the Netherlands, Malta and Serbia. Is family medicine an international discipline? Comparing diagnostic odds ratios across populations. Family practice 2012;**29**(3):299-314 doi: 10.1093/fampra/cmr099[published Online First: Epub Date]|.
31. Soler JK, Okkes I, Wood M, Lamberts H. The coming of age of ICPC: celebrating the 21st birthday of the International Classification of Primary Care. Family practice 2008;**25**(4):312-7 doi: 10.1093/fampra/cmn028[published Online First: Epub Date]|.
32. Minas M, Koukosias N, Zintzaras E, Kostikas K, Gourgoulianis KI. Prevalence of chronic diseases and morbidity in primary health care in central Greece: an epidemiological study. BMC health services research 2010;**10**:252 doi: 10.1186/1472-6963-10-252[published Online First: Epub Date]|.
33. Gershon AS, Wang C, Guan J, Vasilevska-Ristovska J, Cicutto L, To T. Identifying individuals with physcian diagnosed COPD in health administrative databases. Copd 2009;**6**(5):388-94
34. Carnahan RM, Moores KG. Mini-Sentinel's systematic reviews of validated methods for identifying health outcomes using administrative and claims data: methods and lessons learned. Pharmacoepidemiol Drug Saf 2012;**21 Suppl 1**:82-9 doi: 10.1002/pds.2321[published Online First: Epub Date]|.

35. McPheeters ML, Sathe NA, Jerome RN, Carnahan RM. Methods for systematic reviews of administrative database studies capturing health outcomes of interest. Vaccine 2013;**31 Suppl 10**:K2-6 doi: 10.1016/j.vaccine.2013.06.048[published Online First: Epub Date]|.
36. Shamseer L, Moher D, Clarke M, et al. Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015: elaboration and explanation. Bmj 2015;**349**:g7647 doi: 10.1136/bmj.g7647[published Online First: Epub Date]|.
37. Bossuyt PM, Reitsma JB, Bruns DE, et al. Towards complete and accurate reporting of studies of diagnostic accuracy: The STARD Initiative. Annals of internal medicine 2003;**138**(1):40-4
38. West SL, Strom BL, Poole C. *Validity of Pharmacoepidemiologic Drug and Diagnosis Data*: John Wiley & Sons, Ltd, 2007.
39. Reitsma JB, Glas AS, Rutjes AW, Scholten RJ, Bossuyt PM, Zwinderman AH. Bivariate analysis of sensitivity and specificity produces informative summary measures in diagnostic reviews. Journal of clinical epidemiology 2005;**58**(10):982-90 doi: 10.1016/j.jclinepi.2005.02.022[published Online First: Epub Date]|.

Figure . Study screening process (PRISMA Flow Diagram)

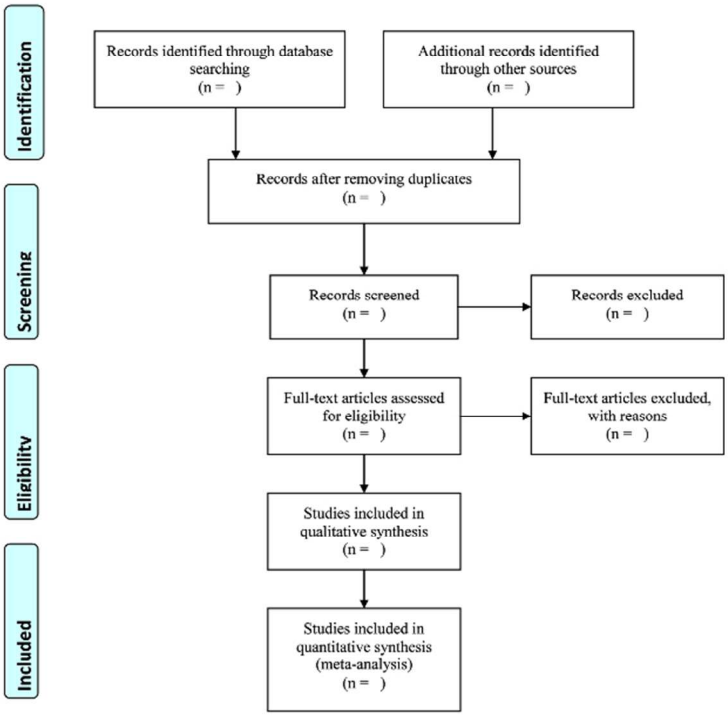

Figure 1. Study screening process
61x86mm (300 x 300 DPI)

Appendix

MEDLINE (via Pubmed) search strategy

1. (health administrative) OR (administrative data) OR (administrative database) OR (claim administrative) OR (International Classification of Diseases) OR "International Classification of Diseases"[Mesh] OR ICD-9-CM OR ICD-10 OR "Database Management Systems"[Mesh] OR "Medical Records Systems, Computerized"[Mesh] OR "CPT" OR "Current procedural terminology"[Mesh] OR (computerized medical records systems) OR (electronic healthcare record) OR (computerized medical record) OR (electronic medical record) OR (automated medical record) OR (electronic patient record) OR CPRD OR GPRD OR Optum OR PHARMO OR HealthCore OR Danish registries
2. (factual databases) OR (geographic information systems) OR (national practitioner data bank) OR (insurance database)
3. #1 OR #2
4. sensitivity or "Sensitivity and Specificity"[Mesh]
5. specificity[Title/Abstract]
6. (positive predictive value) OR (negative predictive value) OR (likelihood ratio) OR (receiver operating characteristic) OR kappa
7. ((case or cases) AND (verificat* OR valid* OR identif* OR definition* OR define* OR evaluat*))
8. Algorithm OR "Algorithm"[Mesh]
9. #4 OR #5 OR #6 OR #7 OR #8
10. emphysema
11. (chronic* bronchitis*)
12. (obstruct* (pulmonary or lung* or airway* or airflow* or bronch* or respirat*))
13. (chronic obstructive pulmonary disease)[Text Word]
14. COPD
15. CAPD
16. "pulmonary disease, chronic obstructive"[MeSH Terms]
17. #10 OR #11 OR #12 OR #13 OR #14 OR #15 OR #16
18. #3 AND #9 AND #17

EMBASE search strategy (via embase.com)

1. health NEAR/3 administrative OR administrative NEAR/3 data OR administrative NEAR/3 database OR claim NEAR/3 administrative OR (International Classification of Diseases) OR 'International Classification of Diseases'/exp OR ICD-9-CM OR ICD-10 OR 'Database Management Systems'/exp OR 'Medical Records Systems, Computerized'/exp OR 'CPT' OR 'Current procedural terminology'/exp OR (computerized medical records systems) OR (electronic healthcare record) OR (computerized medical record) OR (electronic medical record) OR (automated medical record) OR (electronic patient record) OR CPRD OR GPRD OR Optum OR PHARMO OR HealthCore OR Danish registries
2. database:ab,ti OR (('practitioner'/exp OR practitioner) AND data AND bank) OR (('practitioner'/exp OR practitioner) AND ('database'/exp OR database)) OR ('insurance' AND ('database'/exp OR database))
3. #1 OR #2
4. 'sensitivity and specificity'/exp OR 'sensitivity and specificity'
5. specificity:ab,ti
6. 'predictive value of tests'/exp OR 'predictive value of tests'
7. (positive:ab,ti AND predictive:ab,ti AND value:ab,ti) OR (negative:ab,ti AND predictive:ab,ti AND value:ab,ti) OR (likelyhood:ab,ti AND ratio:ab,ti) OR (receiver:ab,ti AND operating:ab,ti AND characteristic:ab,ti) OR kappa:ab,ti
8. case NEAR/1 (verificat* OR valid* OR identif* OR definition* OR define* OR evaluat*)
9. 'algorithms'/exp OR algorithm
10. #4 OR #5 OR #6 OR #7 OR #8 OR #9
11. 'emphysema'/exp
12. 'chronic bronchitis'/exp
13. (obstruct* NEAR/3 (pulmonary or lung* or airway* or airflow* or bronch* or respirat*))
14. 'chronic obstructive pulmonary disease '/exp
15. 'COPD'/exp
16. 'CAPD'/exp
17. #11 OR #12 OR #13 OR #14 OR #15 OR #16
18. #3 AND #10 AND #17

Web of Science search strategy

1. (health NEAR/3 administrative) OR (administrative NEAR/3 data) OR (administrative NEAR/3 database) OR (claim NEAR/3 administrative) OR (International Classification of Diseases) OR ICD-9-CM OR ICD-10 OR (Database Management Systems) OR ("Medical Records Systems" NEAR/2 Computerized) OR "CPT" OR (Current procedural terminology) OR (computerized medical records systems) OR (electronic healthcare record) OR (computerized medical record) OR (electronic medical record) OR (automated medical record) OR (electronic patient record) OR CPRD OR GPRD OR Optum OR PHARMO OR HealthCore OR Danish registries
2. (factual databases) OR (geographic information systems) OR (national practitioner data bank) OR (insurance database)
3. #1 OR #2
4. sensitivity or "Sensitivity and Specificity"
5. specificity
6. (positive predictive value) OR (negative predictive value) OR (likelihood ratio) OR (receiver operating characteristic) OR kappa
7. ((case or cases) AND (verificat* OR valid* OR identif* OR definition* OR define* OR evaluat*))
8. algorithm
9. #4 OR #5 OR #6 OR #7 OR #8
10. emphysema
11. (chronic* NEAR/3 bronchitis*)
12. (obstruct* NEAR/3 (pulmonary or lung* or airway* or airflow* or bronch* or respirat*))
13. chronic obstructive pulmonary disease
14. (COPD)
15. (CAPD)
16. #10 OR #11 OR #12 OR #13 OR #14 OR #15
17. # 3 AND #9 AND #16

The Cochrane Library

1. (health near/3 administrative) or (administrative near/3 data) or (administrative near/3 database) or (claim near/3 administrative) or (International Classification of Diseases) or [mh "International Classification of Diseases"] or ICD-9-CM or ICD-10 or [mh "Database Management Systems"] or [mh "Medical Records Systems, Computerized"] or "CPT" or [mh "Current procedural terminology"] or (computerized medical records systems) or (electronic healthcare record) or (computerized medical record) or (electronic medical record) or (automated medical record) or (electronic patient record) or CPRD or GPRD or Optum or PHARMO or HealthCore or Danish registries
2. (factual databases) OR (geographic information systems) OR (national practitioner data bank) OR (insurance database)
3. #1 OR #2
4. sensitivity or [mh "Sensitivity and Specificity"]
5. specificity:ti,ab,kw
6. (positive predictive value) OR (negative predictive value) OR (likelihood ratio) OR (receiver operating characteristic) OR kappa
7. ((case or cases) AND (verificat* OR valid* OR identif* OR definition* OR define* OR evaluat*))
8. Algorithm OR [mh "Algorithm"]
9. #4 OR #5 OR#6 OR #7 OR #8
10. emphysema
11. (chronic* near/3 bronchitis*)
12. (obstruct* near/3 (pulmonary or lung* or airway* or airflow* or bronch* or respirat*))
13. (chronic obstructive pulmonary disease)
14. COPD
15. CAPD
16. [mh "pulmonary disease, chronic obstructive"]
17. #10 OR #11 OR #12 OR #13 OR #14 OR #15 OR #16
18. # 3 AND #9 AND #17

Appendix 2

Checklist of reporting criteria for studies validating health administrative data algorithms (developed by Benchimol et al., based on the criteria published by the Standards for Reporting of Diagnostic accuracy (STARD) initiative for the accurate reporting of studies using diagnostic studies.

	YES	NO	UNCERTAIN	NOT APPLICABLE
TITLE, KEYWORDS, ABSTRACT				
Identify article as study of assessing diagnostic accuracy				
Identify article as study of administrative data				
INTRODUCTION:				
State disease identification & validation one of goals of study				
METHODS:				
Participants in validation cohort:				
Describe validation cohort (Cohort of patients to which reference standard was applied)				
• Age				
• Disease				
• Severity				
• Location/Jurisdiction				
Describe recruitment procedure of validation cohort				
• Inclusion criteria				
• Exclusion criteria				
Describe patient sampling (random, consecutive, all, etc.)				
Describe data collection				
• Who identified patients and did selection adhere to patient recruitment criteria				
• Who collected data				
• *A priori* data collection form				
• Disease classification				
• Split sample (i.e. re-validation using a separate cohort) a) Training set b) Testing set				
Test Methods:				
Describe number, training and expertise of persons reading reference standard				
If >1 person reading reference standard, quote measure of consistency (e.g. kappa)				

Blinding of interpreters of reference standard to results of classification by administrative data e.g. Chart abstractor blinded to how that chart was coded				
Statistical Methods:				
Describe methods of calculating/comparing diagnostic accuracy				
RESULTS:				
Participants:				
Report when study done, start/end dates of enrollment				
Describe number of people who satisfied inclusion/exclusion criteria				
Study flow diagram				
Test results:				
Report distribution of disease severity				
Report cross-tabulation of index tests by results of reference standard				
Estimates:				
Report at least 4 estimates of diagnostic accuracy				
Diagnostic Accuracy Measures Reported:				
• Sensitivity				
• Spec				
• PPV				
• NPV				
• Likelihood ratios				
• Kappa				
• Area under the ROC curve / c-statistic				
• Accuracy/agreement				
• Other (specify)				
Report accuracy for subgroups (e.g. age, geography, different sex, etc.)				
If PPV/NPV reported, ratio of cases/controls of validation cohort approximate prevalence of condition in the population				
Report 95% confidence intervals for each diagnostic measure				
DISCUSSION:				
Discuss the applicability of the validation findings				

PRISMA-P (Preferred Reporting Items for Systematic review and Meta-Analysis Protocols) 2015 checklist: recommended items to address in a systematic review protocol*

Section and topic	Item No	Checklist item	Page
ADMINISTRATIVE INFORMATION			
Title:			
Identification	1a	Identify the report as a protocol of a systematic review	Page 1
Update	1b	If the protocol is for an update of a previous systematic review, identify as such	
Registration	2	If registered, provide the name of the registry (such as PROSPERO) and registration number	Page 2: Trial registration number PROSPERO 2015 CRD42015029204
Authors:			
Contact	3a	Provide name, institutional affiliation, e-mail address of all protocol authors; provide physical mailing address of corresponding author	Page 1
Contributions	3b	Describe contributions of protocol authors and identify the guarantor of the review	Page 10
Amendments	4	If the protocol represents an amendment of a previously completed or published protocol, identify as such and list changes; otherwise, state plan for documenting important protocol amendments	At this stage there are no relevant amendments to perform
Support:			
Sources	5a	Indicate sources of financial or other support for the review	Page 10 *not funded*
Sponsor	5b	Provide name for the review funder and/or sponsor	Page 10
Role of sponsor or funder	5c	Describe roles of funder(s), sponsor(s), and/or institution(s), if any, in developing the protocol	Page 10
INTRODUCTION			
Rationale	6	Describe the rationale for the review in the context of what is already known	Page 4 and 5
Objectives	7	Provide an explicit statement of the question(s) the review will address with reference to participants, interventions, comparators, and outcomes (PICO)	Page 5
METHODS			
Eligibility criteria	8	Specify the study characteristics (such as PICO, study design, setting,	Page 5, 6:

		time frame) and report characteristics (such as years considered, language, publication status) to be used as criteria for eligibility for the review	
Information sources	9	Describe all intended information sources (such as electronic databases, contact with study authors, trial registers or other grey literature sources) with planned dates of coverage	Page 6.
Search strategy	10	Present draft of search strategy to be used for at least one electronic database, including planned limits, such that it could be repeated	Appendix 1 in Supplemental file
Study records:			Page 7:
Data management	11a	Describe the mechanism(s) that will be used to manage records and data throughout the review	Page 7:.
Selection process	11b	State the process that will be used for selecting studies (such as two independent reviewers) through each phase of the review (that is, screening, eligibility and inclusion in meta-analysis)	Page 7:
Data collection process	11c	Describe planned method of extracting data from reports (such as piloting forms, done independently, in duplicate), any processes for obtaining and confirming data from investigators	Page 7:
Data items	12	List and define all variables for which data will be sought (such as PICO items, funding sources), any pre-planned data assumptions and simplifications	Page 7/8 .
Outcomes and prioritization	13	List and define all outcomes for which data will be sought, including prioritization of main and additional outcomes, with rationale	Page 5
Risk of bias in individual studies	14	Describe anticipated methods for assessing risk of bias of individual studies, including whether this will be done at the outcome or study level, or both; state how this information will be used in data synthesis	Not applicable. The present review will apply the STARD criteria (Appendix 2 in Supplemental file).
Data synthesis	15a	Describe criteria under which study data will be quantitatively synthesised	No cumulative evidence will be presented.
	15b	If data are appropriate for quantitative synthesis, describe planned summary measures, methods of handling data and methods of combining data from studies, including any planned exploration of consistency (such as I^2, Kendall's τ)	
	15c	Describe any proposed additional analyses (such as sensitivity or subgroup analyses, meta-regression)	
	15d	If quantitative synthesis is not appropriate, describe the type of summary planned	
Meta-bias(es)	16	Specify any planned assessment of meta-bias(es) (such as publication	Not applicable

		bias across studies, selective reporting within studies)	
Confidence in cumulative evidence	17	Describe how the strength of the body of evidence will be assessed (such as GRADE)	The present review will apply the STARD criteria. Page 8 . (Appendix 2 in Supplemental file).

*** It is strongly recommended that this checklist be read in conjunction with the PRISMA-P Explanation and Elaboration (cite when available) for important clarification on the items. Amendments to a review protocol should be tracked and dated. The copyright for PRISMA-P (including checklist) is held by the PRISMA-P Group and is distributed under a Creative Commons Attribution Licence 4.0.**

From: Shamseer L, Moher D, Clarke M, Ghersi D, Liberati A, Petticrew M, Shekelle P, Stewart L, PRISMA-P Group. Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015: elaboration and explanation. BMJ. 2015 Jan 2;349(jan02 1):g7647.

BMJ Open

Validation of Chronic Obstructive Pulmonary Disease (COPD) Diagnoses in Healthcare Databases: A Systematic Review Protocol

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2016-011777.R2
Article Type:	Protocol
Date Submitted by the Author:	10-May-2016
Complete List of Authors:	Rimland, Joseph; Italian National Research Center on Aging (INRCA), Geriatrics and Geriatric Emergency Care Abraha, Iosief; Regional Health Authority of Umbria, Health Planning Service Luchetta, Maria Laura; Azienda USL Umbria 1, General Medicine Cozzolino, Francesco; Regional Health Authority of Umbria, Health Planning Service Orso, Massimiliano; Regional Health Authority of Umbria, Health Planning Service Cherubini, Antonio; Italian National Research Center on Aging (INRCA), Geriatrics and Geriatric Emergency Care Dell'Aquila, Giuseppina; Italian National Research Center on Aging (INRCA), Geriatrics and Geriatric Emergency Care Chiatti, Carlos; Italian National Research Centre on Aging (INRCA), Italy Ambrosio, Giuseppe; University of Perugia School of Medicine, Cardiology; Ospedale S. Maria della Misericordia, Medical Administration Montedori, Alessandro; Regional Health Authority of Umbria, Health Planning Service
Primary Subject Heading:	Research methods
Secondary Subject Heading:	Respiratory medicine, Public health, Epidemiology
Keywords:	Administrative database, Chronic pulmonary obstructive disease, sensitivity and specificity, systematic review, Electronic Health Record, COPD algorithm

Validation of Chronic Obstructive Pulmonary Disease (COPD) Diagnoses in Healthcare Databases: A Systematic Review Protocol

Joseph M Rimland, Iosief Abraha, Maria Laura Luchetta, Francesco Cozzolino, Massimiliano Orso, Antonio Cherubini, Giuseppina Dell'Aquila, Carlos Chiatti, Giuseppe Ambrosio, Alessandro Montedori.

Author affiliations:
Health Planning Service, Regional Health Authority of Umbria, Perugia, Italy
Iosief Abraha
Alessandro Montedori
Francesco Cozzolino
Massimiliano Orso

Scientific Directorate, Italian National Research Center on Aging, Ancona, Italy
Carlos Chiatti

Geriatrics and Geriatric Emergency Care, Italian National Research Center on Aging, Ancona, Italy
Antonio Cherubini
Joseph M Rimland
Giuseppina Dell'Aquila

Azienda USL Umbria 1, General Medicine, Perugia, Italy
Maria Laura Luchetta

University of Perugia School of Medicine, Cardiology, Perugia, Italy
Giuseppe Ambrosio

Correspondence to:
Dr. Iosief Abraha
Health Planning Service
Regional Health Authority of Umbria
Via Mario Angeloni, 61
06124 Perugia (Italy)
tel +39 075 504 5251
cell. +39349 077 0910
fax +39 075 504 5569
e-mail: iosief_a@yahoo.it
iabraha@regione.umbria.it

Abstract

Introduction Healthcare databases are useful sources to investigate the epidemiology of chronic obstructive pulmonary disease (COPD), to assess longitudinal outcomes in subjects with COPD, and to develop disease management strategies. However, in order to constitute a reliable source for research, healthcare databases need to be validated. The aim of this protocol is to perform the first systematic review of studies reporting the validation of codes related to COPD diagnoses in healthcare databases.

Methods and analysis MEDLINE, EMBASE, Web of Science and the Cochrane Library databases will be searched, using appropriate search strategies. Studies that evaluated the validity of COPD codes (such as the *International Classification of Diseases 9th Revision* and *10th Revision* system; the *Real codes system* or the *International Classification of Primary Care*) in healthcare databases will be included. Inclusion criteria will be: (a) the presence of a reference standard case definition for COPD; (b) the presence of at least one test measure (e.g. sensitivity, positive predictive values, etc.); and (c) the use of a healthcare database (including administrative claims databases, electronic healthcare databases or COPD registries) as a data source. Pairs of reviewers will independently abstract data using standardised forms and will assess quality using a checklist based on the Standards for Reporting of Diagnostic accuracy (STARD) criteria. This systematic review protocol has been produced in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses Protocol (PRISMA-P) 2015 statement.

Ethics and dissemination Ethics approval is not required. Results of this study will be submitted to a peer-reviewed journal for publication. The results from this systematic review will be used for outcome research on COPD and will serve as a guide to identify appropriate case definitions of COPD, and reference standards, for researchers involved in validating healthcare databases.

Trial registration number PROSPERO 2015 CRD42015029204

Strengths and limitations of this study

- Validation of diagnosis codes for Chronic Obstructive Pulmonary Disease (COPD) using healthcare databases can contribute to health outcome research. The diagnosis codes may include the *International Classification of Diseases 9th Revision* and *10th Revision* (ICD-9; ICD-10) system, the *Real code system* and the *International Classification of Primary Care* system.
- This review will be the first to systematically identify and evaluate primary studies that validated the accuracy of healthcare databases with ICD-9 and ICD-10 codes for COPD
- It is expected that different healthcare databases validate different algorithms to identify COPD resulting in important heterogeneity. Validated algorithms are context specific and may not be generalizable to other settings.

Introduction

Chronic Obstructive Pulmonary Disease (COPD) is a global health problem [1 2]. It is distinguished by continuous airflow restriction, is frequently progressive and is associated with a chronically increased airway and lung inflammatory reaction to gases or particles [3 4]. COPD is correlated with significant morbidity and mortality and is the fourth leading cause of death worldwide [5]. On the basis of WHO estimates (2004), 64 million people had moderate to severe COPD, which led to 3 million deaths [6]. The burden of COPD is estimated to increase in the near future, because of continued exposure to risk factors and ageing of the population [3 4]. Smoking is the main cause of COPD, but other factors, especially exposure to occupational or environmental airborne irritants, may also contribute to the development of this group of lung diseases [3 4].

Healthcare databases are increasingly being used to examine features of health care delivery, including practice patterns, quality of care, safety and efficacy of drugs, and epidemiological studies. Some of the advantages of healthcare databases included the minimisation of recall bias, better generalizability than randomised trials and better cost-effectiveness approach to research compared to primary data collection[7].To be reliably used for research, healthcare databases need to be validated concerning the disease of interest[8-12]. This means that the content of the databases (e.g., a code of a disease) need to be ascertained using a reference standard (e.g., medical chart)[13]. Alternatively, algorithms can be developed by combining multiple codes – or sets of codes (e.g., diagnosis codes plus prescription or spirometry data) to enhance the ability to identify events of interest in the database [13-17].

Healthcare databases generally encompass administrative claims data and Electronic Health Records (EHR). Administrative claims databases routinely collect data passively, for administrative purposes, for health services delivered by healthcare providers and facilities [18]. The patient information collected includes demographics (name, address, birthdate, gender, and marital status), the dates of healthcare services delivered and charges for the services, diagnostic procedures

performed and healthcare service provider information and in some occasions employment, insurance status, occupational limitations.

Administrative claims databases are excellent resources to investigate the epidemiology [17 19 20], and the burden of COPD [21 22] and to evaluate longitudinal outcomes of a disease [23 24]. Results from analysing these databases can assist in developing disease management strategies (including education regarding the disease, optimisation of evidence-based medications, information, case manager support and institution of self-management principles) to improve the health of subjects suffering from COPD [25].

Electronic Health Records (EHRs) consist of digital files used by healthcare providers for patient care, and unlike administrative claims databases, include clinical notes, medical records, the treatment histories of patients, and prescription records, as well as radiology and laboratory data[26]. Despite EHRs are not established for research purposes, similar to most administrative databases, they are frequently used for healthcare delivery and facilitation of decision-making processes as well as research [26 27].

The Clinical Practice Research Datalink (CPRD), used in the UK, is one such EHR. It is an excellent resource in which to study COPD, as it is based on a large cohort, contains disease severity indicators and long-term follow-up information from a patient's integrated medical history [28-30].

Generally, administrative claims databases use the *International Classification of Diseases, 9th Revision* (ICD-9) codes for COPD (491, 492 or 496), or the *International Classification of Diseases, 10th Revision* (ICD-10) codes (J43 and J44). EHRs such as the UK CPRD database employs the Read code, which is a hierarchical clinical coding system of medical and prescription terms [28]. Some Read codes for COPD are 1001, 9876 and 10863. (See [28] for a list of COPD-related Read codes). The International Classification of Primary Care (ICPC) is another coding system, which is widely used in primary health care and in research[31-33]. The codes for COPD in the ICPC system are R79 and R95.

There are several studies that assessed the validity of healthcare databases for COPD [13 17 28 34], however, to our knowledge, no systematic assessment of algorithms or case definitions of COPD have been published in the medical literature. With the present protocol, we aim to systematically evaluate validation studies of diagnostic codes or algorithms to identify cases of COPD.

Research question

The primary research question is the accuracy of algorithms to correctly identify patients with COPD in healthcare databases (administrative claims, EHR, or COPD registries). The target populations are patients with COPD, the index test will be healthcare data algorithms for COPD, the reference standard will be medical charts, validated electronic health records or COPD registries. Our primary outcome is the accuracy (expressed in terms of sensitivity, specificity and positive and negative predictive values) of healthcare data algorithms to discriminate cases of COPD.

Methods

Literature search

Comprehensive searches of MEDLINE, EMBASE, the Web of Science and the Cochrane Library from their inception, will be performed to identify published peer-reviewed articles. A search strategy will be employed that we developed based on the combination of: (a) keywords and MeSH terms to identify records concerning COPD; and (b) a search strategy, based on the combination of terms used by Benchimol et al. [18], the Mini-Sentinel's program [35 36], and a systematic review that evaluated EHR based primary studies[26]. The developed search strategy is available as supplementary material (**Appendix**). To retrieve additional articles, relevant reference lists of key articles will be hand searched. The "Cited-By" tools in PubMed and Google Scholar will also be used to find relevant articles that cited the article of interest, identified through the aforementioned search strategy. Titles and abstracts will be screened for eligibility by two independent reviewers and discrepancies will be resolved by discussion.

This review protocol has been prepared according to the Preferred Reporting Items for Systematic reviews and Meta-Analysis Protocols (PRISMA-P) 2015 Statement [37] and the results will be presented following the PRISMA flow diagram (**Figure**). This protocol has also been published in the PROSPERO International Prospective Register of Systematic Reviews with registration number CRD42015029204 (http://www.crd.york.ac.uk/PROSPERO).

Inclusion criteria

Full-texts of eligible peer-reviewed articles, without limits on publication date and published in English, that used healthcare data to validate diagnosis codes for COPD diagnoses, will be obtained. For each study, the following inclusion criteria will be applied: (a) the presence of a reference standard case definition for the disease of interest; (b) the presence of at least one test measure (e.g., sensitivity, positive predictive values, etc.); (c) the use of an administrative claims or EHR database

as a data source; and (d) the use of a database from a representative sample of the general population[15] [26].

At the initial stage, titles and abstracts will be screened for potentially eligible studies. Subsequently, full texts of articles will be obtained and assessed to determine if they meet the inclusion and exclusion criteria. Data abstraction will be conducted using standardised data collection forms, which will first be tested on a sample of eligible articles. Two review authors working independently, and in duplicate, will carry out title, abstract and full-text screening and data abstraction. Any discrepancies will be resolved by consensus, and where necessary, a third review author will be involved. Calibration exercises will be performed at each level of the process.

Data extraction

Data extraction will include the following information:

(a) the details of the included study (containing the title, the year of publication and the journal, the country of origin, and the sources of funding; the first author will be used as the study ID);

(b) the disease of interest (COPD);

(c) the code tested (such as ICD-9, ICD-10, or R79 and R95);

(d) the algorithm(s) tested including COPD code, prescription fills (e.g., bronchodilators), use of spirometry, Current Procedural Terminology, timing of diagnosis, etc.;

(e) any information about the performance of the COPD definition/algorithm in sub-populations (e.g. age group, sex, smoking status, GOLD grade of airflow limitation[2], socioeconomic status, WHO Body Mass Index (BMI) category, previous record of asthma diagnosis[28])

(f) the target population from which the healthcare data were collected;

(g) the type of healthcare database used (e.g., hospitalisation discharge data, electronic health record etc.);

(h) the modality of algorithm development (e.g., using Classification and Regression Trees, logistic regression, expert opinion…);

(i) external validation;

(j) the use of training and testing cohorts;

(k) the reference standard used to determine the validity of the diagnostic code (e.g., medical chart review, patient self-reports, disease registry, etc.);

(l) the characteristic of the test used to determine the validity of the diagnostic code or algorithm (e.g., sensitivity, specificity, positive predictive values (PPVs) and negative predictive values (NPVs), area under the receiver operating characteristic curve, likelihood ratios, and kappa statistics);

Quality assessment

The design and methods of the included primary studies will be assessed using a checklist developed by Benchimol et al. [18], based on the criteria published by the Standards for Reporting of Diagnostic accuracy (STARD) initiative for the accurate reporting of diagnostic studies [38]. This standardized checklist is composed of 40 items to assess the quality of the methods and the reporting of studies that validated codes or algorithms used to identify patients with the disease of interest within a healthcare database (**Appendix**). Two reviewers will be involved in the quality assessment and will work in duplicate and independently. Any disagreement will be solved by discussion. The presence of potential biases within the studies will be reported descriptively.

No subgroup analysis or publication bias assessment are anticipated.

Analysis

For each algorithm, the performance statistics, provided in each of the included studies, will be abstracted. Validation statistics may include sensitivity, specificity, PPV and NPV. Sensitivity measures the degree to which a diagnosis code (e.g., ICD-9 491 or Read code 1001) correctly

identifies individuals possessing the characteristic of interest (i.e., COPD) in the source used as a reference standard (e.g. medical chart) [39]. PPV is the number of true positives divided by the total number of cases receiving the code and expresses the likelihood that the code corresponds to a true-positive case. NPV is the number of true negatives divided by the total number of cases without the code of interest and expresses the likelihood that the absence of the code corresponds to a true-negative case. Where possible, PPVs and NPVs will be calculated if not reported. Ninety-five percent confidence intervals (95% CI) will be calculated when they are not reported in the articles. Where possible, validation statistics will be aggregated and stratified by healthcare data source (outpatient vs. inpatient data), type of EHR code (ICD-9, ICD-10, Read, etc.), and country of origin.

Meta-analysis

Where there are studies with homogeneous data, we will use raw data to construct meta-analyses. A bivariate model will be used to derive summary estimates of sensitivity and specificity and their 95% CIs[40]. Data will be analysed using a random-effects model so that sensitivity and specificity are assumed to vary across studies. In addition, summary receiver operating characteristic (ROC) curves will be constructed and pooled estimates of LR+, LR-, and diagnostic odds ratios will be calculated. Heterogeneity will be assessed by visual inspection of forest plots and ROC plots as well as regression analysis suggested by Reitsma [40]. Where there is important heterogeneity we will not pool the data.

Ethics and dissemination

This review protocol will use publicly available data without directly involving human participants, hence approval from an ethics committee is not required. An outline of the protocol has been published in the PROSPERO International Prospective Register of Systematic Reviews in 2015, registration number CRD42015029204. The results will summarise the studies that validated diagnostic codes for Chronic Obstructive Pulmonary Disease in administrative data. Where possible, a quantitative synthesis of the accuracy data will be provided and the outcomes using different algorithms will be discussed. Findings of the review will be presented at relevant scientific conferences and disseminated through publication in a peer-reviewed journal.

Footnotes

Contributors IA, JMR, and MLL conceived the study. JMR, IA, MLL, FC, MO, AC, GD, CC, GA, and AM were responsible for designing the protocol. MLL, JMR and IA drafted the protocol manuscript. JMR, IA, FC, and MO developed the search strategy. JMR, IA, MLL, FC, MO, AC, GD, CC, GA, and AM critically revised the successive versions of the manuscript and approved the final version. IA acts as guarantor.

Funding Currently the study is not funded.

Competing interests None.

References

1. GOLD GIfCLO. Pocket guide to COPD diagnosis, management and prevention. www.goldcopd.com. www.goldcopd.com ed, 2015.
2. Vestbo J, Hurd SS, Agusti AG, et al. Global strategy for the diagnosis, management, and prevention of chronic obstructive pulmonary disease: GOLD executive summary. American journal of respiratory and critical care medicine 2013;**187**(4):347-65 doi: 10.1164/rccm.201204-0596PP[published Online First: Epub Date]|.
3. Jones PW, Nadeau G, Small M, Adamek L. Characteristics of a COPD population categorised using the GOLD framework by health status and exacerbations. Respiratory Medicine 2014;**108**(1):129-35 doi: http://dx.doi.org/10.1016/j.rmed.2013.08.015[published Online First: Epub Date]|.
4. Rabe KF, Hurd S, Anzueto A, et al. Global Strategy for the Diagnosis, Management, and Prevention of Chronic Obstructive Pulmonary Disease. American Journal of Respiratory and Critical Care Medicine 2007;**176**(6):532-55 doi: 10.1164/rccm.200703-456SO[published Online First: Epub Date]|.
5. Lange P, Marott JL, Vestbo J, et al. Prediction of the Clinical Course of Chronic Obstructive Pulmonary Disease, Using the New GOLD Classification. American Journal of Respiratory and Critical Care Medicine 2012;**186**(10):975-81 doi: 10.1164/rccm.201207-1299OC[published Online First: Epub Date]|.
6. Adeloye D, Chua S, Lee C, et al. Global and regional estimates of COPD prevalence: Systematic review and meta-analysis. Journal of global health 2015;**5**(2):020415 doi: 10.7189/jogh.05-020415[published Online First: Epub Date]|.
7. Schneeweiss S, Avorn J. A review of uses of health care utilization databases for epidemiologic research on therapeutics. Journal of clinical epidemiology 2005;**58**(4):323-37 doi: 10.1016/j.jclinepi.2004.10.012[published Online First: Epub Date]|.
8. Abraha I, Giovannini G, Serraino D, Fusco M, Montedori A. Validity of Breast, Lung and Colorectal Cancer Diagnoses in Administrative Databases: A Systematic Review Protocol. BMJ Open 2016:010409
9. Abraha I, Serraino D, Giovannini G, et al. Validity of ICD-9-CM Codes for Breast, Lung, and Colorectal Cancers in Three Italian Administrative Healthcare Databases: A Diagnostic Accuracy Study Protocol BMJ Open 2016;**Accept (22-Feb-2016)**:010547
10. West SL, Ritchey ME, Poole C. Validity of Pharmacoepidemiologic Drug and Diagnosis Data. Pharmacoepidemiology: Wiley-Blackwell, 2012:757-94.
11. Abraha I, Orso M, Grilli P, et al. The Current State of Validation of Administrative Healthcare Databases in Italy: A Systematic Review. International Journal of Statistics in Medical Research 2014;**3**(3):309-20
12. Abraha I, Montedori A, Eusebi P, et al. The Current State of Validation of Administrative Healthcare Databases in Italy: A Systematic Review. Pharmacoepidemiology and Drug Safety 2012;**21**:400-00
13. Lacasse Y, Daigle JM, Martin S, Maltais F. Validity of chronic obstructive pulmonary disease diagnoses in a large administrative database. Canadian respiratory journal : journal of the Canadian Thoracic Society 2012;**19**(2):e5-9
14. Lix LM, De Coster C, Currie R. *Defining and validating chronic diseases: an administrative data approach*: Manitoba Centre for Health Policy Winnipeg, 2006.
15. Chubak J, Pocobelli G, Weiss NS. Tradeoffs between accuracy measures for electronic health care data algorithms. Journal of clinical epidemiology 2012;**65**(3):343-49.e2 doi: 10.1016/j.jclinepi.2011.09.002[published Online First: Epub Date]|.
16. Sharifi M, Krishanswami S, McPheeters ML. A systematic review of validated methods to capture acute bronchospasm using administrative or claims data. Vaccine 2013;**31 Suppl 10**:K12-20 doi: 10.1016/j.vaccine.2013.06.091[published Online First: Epub Date]|.
17. Green ME, Natajaran N, O'Donnell DE, et al. Chronic obstructive pulmonary disease in primary care: an epidemiologic cohort study from the Canadian Primary Care Sentinel Surveillance Network. CMAJ open 2015;**3**(1):E15-E22 doi: 10.9778/cmajo.20140040[published Online First: Epub Date]|.
18. Benchimol EI, Manuel DG, To T, Griffiths AM, Rabeneck L, Guttmann A. Development and use of reporting guidelines for assessing the quality of validation studies of health administrative data.

Journal of clinical epidemiology 2011;**64**(8):821-9 doi: 10.1016/j.jclinepi.2010.10.006[published Online First: Epub Date]|.
19. Anecchino C, Rossi E, Fanizza C, De Rosa M, Tognoni G, Romero M. Prevalence of chronic obstructive pulmonary disease and pattern of comorbidities in a general population. International Journal of Chronic Obstructive Pulmonary Disease 2007;**2**(4):567-74
20. Faustini A, Canova C, Cascini S, et al. The reliability of hospital and pharmaceutical data to assess prevalent cases of chronic obstructive pulmonary disease. Copd 2012;**9**(2):184-96 doi: 10.3109/15412555.2011.654014[published Online First: Epub Date]|.
21. Bustacchini S, Chiatti C, Furneri G, Lattanzio F, Mantovani LG. The economic burden of chronic obstructive pulmonary disease in the elderly: results from a systematic review of the literature. Current opinion in pulmonary medicine 2011;**17 Suppl 1**:S35-41 doi: 10.1097/01.mcp.0000410746.82840.79[published Online First: Epub Date]|.
22. Simoni-Wastila L, Blanchette CM, Qian J, et al. Burden of chronic obstructive pulmonary disease in Medicare beneficiaries residing in long-term care facilities. The American journal of geriatric pharmacotherapy 2009;**7**(5):262-70 doi: 10.1016/j.amjopharm.2009.11.003[published Online First: Epub Date]|.
23. Ismaila A, Corriveau D, Vaillancourt J, et al. Impact of adherence to treatment with tiotropium and fluticasone propionate/salmeterol in chronic obstructive pulmonary diseases patients. Current medical research and opinion 2014;**30**(7):1427-36 doi: 10.1185/03007995.2014.908828[published Online First: Epub Date]|.
24. Abudagga A, Sun SX, Tan H, Solem CT. Exacerbations among chronic bronchitis patients treated with maintenance medications from a US managed care population: an administrative claims data analysis. Int J Chron Obstruct Pulmon Dis 2013;**8**:175-85 doi: 10.2147/copd.s40437[published Online First: Epub Date]|.
25. Rice KL, Dewan N, Bloomfield HE, et al. Disease management program for chronic obstructive pulmonary disease: a randomized controlled trial. American journal of respiratory and critical care medicine 2010;**182**(7):890-6 doi: 10.1164/rccm.200910-1579OC[published Online First: Epub Date]|.
26. Dean BB, Lam J, Natoli JL, Butler Q, Aguilar D, Nordyke RJ. Review: Use of Electronic Medical Records for Health Outcomes Research: A Literature Review. Medical Care Research and Review 2009;**66**(6):611-38 doi: 10.1177/1077558709332440[published Online First: Epub Date]|.
27. Williams F, Boren SA. The role of the electronic medical record (EMR) in care delivery development in developing countries: a systematic review. Informatics in primary care 2008;**16**(2):139-45
28. Quint JK, Mullerova H, DiSantostefano RL, et al. Validation of chronic obstructive pulmonary disease recording in the Clinical Practice Research Datalink (CPRD-GOLD). BMJ Open 2014;**4**(7):e005540 doi: 10.1136/bmjopen-2014-005540[published Online First: Epub Date]|.
29. Rothnie KJ, Mullerova H, Hurst JR, et al. Validation of the Recording of Acute Exacerbations of COPD in UK Primary Care Electronic Healthcare Records. PloS one 2016;**11**(3):e0151357 doi: 10.1371/journal.pone.0151357[published Online First: Epub Date]|.
30. Wurst KE, St Laurent S, Mullerova H, Davis KJ. Characteristics of patients with COPD newly prescribed a long-acting bronchodilator: a retrospective cohort study. Int J Chron Obstruct Pulmon Dis 2014;**9**:1021-31 doi: 10.2147/copd.s58258[published Online First: Epub Date]|.
31. Soler JK, Okkes I, Oskam S, et al. An international comparative family medicine study of the Transition Project data from the Netherlands, Malta and Serbia. Is family medicine an international discipline? Comparing diagnostic odds ratios across populations. Family practice 2012;**29**(3):299-314 doi: 10.1093/fampra/cmr099[published Online First: Epub Date]|.
32. Soler JK, Okkes I, Wood M, Lamberts H. The coming of age of ICPC: celebrating the 21st birthday of the International Classification of Primary Care. Family practice 2008;**25**(4):312-7 doi: 10.1093/fampra/cmn028[published Online First: Epub Date]|.
33. Minas M, Koukosias N, Zintzaras E, Kostikas K, Gourgoulianis KI. Prevalence of chronic diseases and morbidity in primary health care in central Greece: an epidemiological study. BMC health services research 2010;**10**:252 doi: 10.1186/1472-6963-10-252[published Online First: Epub Date]|.
34. Gershon AS, Wang C, Guan J, Vasilevska-Ristovska J, Cicutto L, To T. Identifying individuals with physcian diagnosed COPD in health administrative databases. Copd 2009;**6**(5):388-94
35. Carnahan RM, Moores KG. Mini-Sentinel's systematic reviews of validated methods for identifying health outcomes using administrative and claims data: methods and lessons learned.

Pharmacoepidemiol Drug Saf 2012;**21 Suppl 1**:82-9 doi: 10.1002/pds.2321[published Online First: Epub Date]|.
36. McPheeters ML, Sathe NA, Jerome RN, Carnahan RM. Methods for systematic reviews of administrative database studies capturing health outcomes of interest. Vaccine 2013;**31 Suppl 10**:K2-6 doi: 10.1016/j.vaccine.2013.06.048[published Online First: Epub Date]|.
37. Shamseer L, Moher D, Clarke M, et al. Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015: elaboration and explanation. Bmj 2015;**349**:g7647 doi: 10.1136/bmj.g7647[published Online First: Epub Date]|.
38. Bossuyt PM, Reitsma JB, Bruns DE, et al. Towards complete and accurate reporting of studies of diagnostic accuracy: The STARD Initiative. Annals of internal medicine 2003;**138**(1):40-4
39. West SL, Strom BL, Poole C. *Validity of Pharmacoepidemiologic Drug and Diagnosis Data*: John Wiley & Sons, Ltd, 2007.
40. Reitsma JB, Glas AS, Rutjes AW, Scholten RJ, Bossuyt PM, Zwinderman AH. Bivariate analysis of sensitivity and specificity produces informative summary measures in diagnostic reviews. Journal of clinical epidemiology 2005;**58**(10):982-90 doi: 10.1016/j.jclinepi.2005.02.022[published Online First: Epub Date]|.

Figure . Study screening process (PRISMA Flow Diagram)

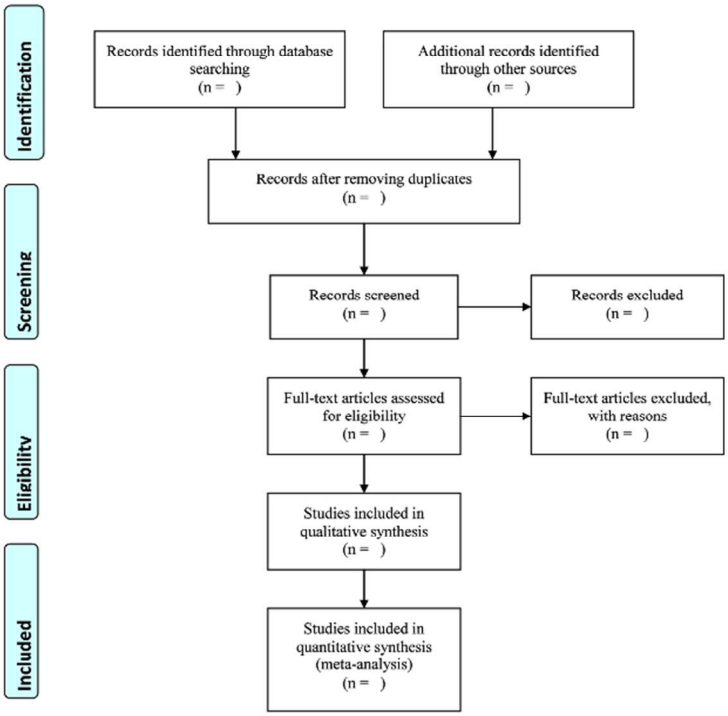

Figure 1. Study screening process
61x86mm (300 x 300 DPI)

Appendix

MEDLINE (via Pubmed) search strategy

1. (health administrative) OR (administrative data) OR (administrative database) OR (claim administrative) OR (International Classification of Diseases) OR "International Classification of Diseases"[Mesh] OR ICD-9-CM OR ICD-10 OR "Database Management Systems"[Mesh] OR "Medical Records Systems, Computerized"[Mesh] OR "CPT" OR "Current procedural terminology"[Mesh] OR (computerized medical records systems) OR (electronic healthcare record) OR (computerized medical record) OR (electronic medical record) OR (automated medical record) OR (electronic patient record) OR CPRD OR GPRD OR Optum OR PHARMO OR HealthCore OR Danish registries
2. (factual databases) OR (geographic information systems) OR (national practitioner data bank) OR (insurance database)
3. #1 OR #2
4. sensitivity or "Sensitivity and Specificity"[Mesh]
5. specificity[Title/Abstract]
6. (positive predictive value) OR (negative predictive value) OR (likelihood ratio) OR (receiver operating characteristic) OR kappa
7. ((case or cases) AND (verificat* OR valid* OR identif* OR definition* OR define* OR evaluat*))
8. Algorithm OR "Algorithm"[Mesh]
9. #4 OR #5 OR #6 OR #7 OR #8
10. emphysema
11. (chronic* bronchitis*)
12. (obstruct* (pulmonary or lung* or airway* or airflow* or bronch* or respirat*))
13. (chronic obstructive pulmonary disease)[Text Word]
14. COPD
15. CAPD
16. "pulmonary disease, chronic obstructive"[MeSH Terms]
17. #10 OR #11 OR #12 OR #13 OR #14 OR #15 OR #16
18. #3 AND #9 AND #17

EMBASE search strategy (via embase.com)

1. health NEAR/3 administrative OR administrative NEAR/3 data OR administrative NEAR/3 database OR claim NEAR/3 administrative OR (International Classification of Diseases) OR 'International Classification of Diseases'/exp OR ICD-9-CM OR ICD-10 OR 'Database Management Systems'/exp OR 'Medical Records Systems, Computerized'/exp OR 'CPT' OR 'Current procedural terminology'/exp OR (computerized medical records systems) OR (electronic healthcare record) OR (computerized medical record) OR (electronic medical record) OR (automated medical record) OR (electronic patient record) OR CPRD OR GPRD OR Optum OR PHARMO OR HealthCore OR Danish registries
2. database:ab,ti OR (('practitioner'/exp OR practitioner) AND data AND bank) OR (('practitioner'/exp OR practitioner) AND ('database'/exp OR database)) OR ('insurance' AND ('database'/exp OR database))
3. #1 OR #2
4. 'sensitivity and specificity'/exp OR 'sensitivity and specificity'
5. specificity:ab,ti
6. 'predictive value of tests'/exp OR 'predictive value of tests'
7. (positive:ab,ti AND predictive:ab,ti AND value:ab,ti) OR (negative:ab,ti AND predictive:ab,ti AND value:ab,ti) OR (likelyhood:ab,ti AND ratio:ab,ti) OR (receiver:ab,ti AND operating:ab,ti AND characteristic:ab,ti) OR kappa:ab,ti
8. case NEAR/1 (verificat* OR valid* OR identif* OR definition* OR define* OR evaluat*)
9. 'algorithms'/exp OR algorithm
10. #4 OR #5 OR #6 OR #7 OR #8 OR #9
11. 'emphysema'/exp
12. 'chronic bronchitis'/exp
13. (obstruct* NEAR/3 (pulmonary or lung* or airway* or airflow* or bronch* or respirat*))
14. 'chronic obstructive pulmonary disease '/exp
15. 'COPD'/exp
16. 'CAPD'/exp
17. #11 OR #12 OR #13 OR #14 OR #15 OR #16
18. #3 AND #10 AND #17

Web of Science search strategy

1. (health NEAR/3 administrative) OR (administrative NEAR/3 data) OR (administrative NEAR/3 database) OR (claim NEAR/3 administrative) OR (International Classification of Diseases) OR ICD-9-CM OR ICD-10 OR (Database Management Systems) OR ("Medical Records Systems" NEAR/2 Computerized) OR "CPT" OR (Current procedural terminology) OR (computerized medical records systems) OR (electronic healthcare record) OR (computerized medical record) OR (electronic medical record) OR (automated medical record) OR (electronic patient record) OR CPRD OR GPRD OR Optum OR PHARMO OR HealthCore OR Danish registries
2. (factual databases) OR (geographic information systems) OR (national practitioner data bank) OR (insurance database)
3. #1 OR #2
4. sensitivity or "Sensitivity and Specificity"
5. specificity
6. (positive predictive value) OR (negative predictive value) OR (likelihood ratio) OR (receiver operating characteristic) OR kappa
7. ((case or cases) AND (verificat* OR valid* OR identif* OR definition* OR define* OR evaluat*))
8. algorithm
9. #4 OR #5 OR #6 OR #7 OR #8
10. emphysema
11. (chronic* NEAR/3 bronchitis*)
12. (obstruct* NEAR/3 (pulmonary or lung* or airway* or airflow* or bronch* or respirat*))
13. chronic obstructive pulmonary disease
14. (COPD)
15. (CAPD)
16. #10 OR #11 OR #12 OR #13 OR #14 OR #15
17. # 3 AND #9 AND #16

The Cochrane Library

1. (health near/3 administrative) or (administrative near/3 data) or (administrative near/3 database) or (claim near/3 administrative) or (International Classification of Diseases) or [mh "International Classification of Diseases"] or ICD-9-CM or ICD-10 or [mh "Database Management Systems"] or [mh "Medical Records Systems, Computerized"] or "CPT" or [mh "Current procedural terminology"] or (computerized medical records systems) or (electronic healthcare record) or (computerized medical record) or (electronic medical record) or (automated medical record) or (electronic patient record) or CPRD or GPRD or Optum or PHARMO or HealthCore or Danish registries
2. (factual databases) OR (geographic information systems) OR (national practitioner data bank) OR (insurance database)
3. #1 OR #2
4. sensitivity or [mh "Sensitivity and Specificity"]
5. specificity:ti,ab,kw
6. (positive predictive value) OR (negative predictive value) OR (likelihood ratio) OR (receiver operating characteristic) OR kappa
7. ((case or cases) AND (verificat* OR valid* OR identif* OR definition* OR define* OR evaluat*))
8. Algorithm OR [mh "Algorithm"]
9. #4 OR #5 OR#6 OR #7 OR #8
10. emphysema
11. (chronic* near/3 bronchitis*)
12. (obstruct* near/3 (pulmonary or lung* or airway* or airflow* or bronch* or respirat*))
13. (chronic obstructive pulmonary disease)
14. COPD
15. CAPD
16. [mh "pulmonary disease, chronic obstructive"]
17. #10 OR #11 OR #12 OR #13 OR #14 OR #15 OR #16
18. # 3 AND #9 AND #17

Appendix 2

Checklist of reporting criteria for studies validating health administrative data algorithms (developed by Benchimol et al., based on the criteria published by the Standards for Reporting of Diagnostic accuracy (STARD) initiative for the accurate reporting of studies using diagnostic studies.

	YES	NO	UNCERTAIN	NOT APPLICABLE
TITLE, KEYWORDS, ABSTRACT				
Identify article as study of assessing diagnostic accuracy				
Identify article as study of administrative data				
INTRODUCTION:				
State disease identification & validation one of goals of study				
METHODS:				
Participants in validation cohort:				
Describe validation cohort (Cohort of patients to which reference standard was applied)				
• Age				
• Disease				
• Severity				
• Location/Jurisdiction				
Describe recruitment procedure of validation cohort				
• Inclusion criteria				
• Exclusion criteria				
Describe patient sampling (random, consecutive, all, etc.)				
Describe data collection				
• Who identified patients and did selection adhere to patient recruitment criteria				
• Who collected data				
• *A priori* data collection form				
• Disease classification				
• Split sample (i.e. re-validation using a separate cohort) a) Training set b) Testing set				
Test Methods:				
Describe number, training and expertise of persons reading reference standard				
If >1 person reading reference standard, quote measure of consistency (e.g. kappa)				

Blinding of interpreters of reference standard to results of classification by administrative data e.g. Chart abstractor blinded to how that chart was coded				
Statistical Methods:				
Describe methods of calculating/comparing diagnostic accuracy				
RESULTS:				
Participants:				
Report when study done, start/end dates of enrollment				
Describe number of people who satisfied inclusion/exclusion criteria				
Study flow diagram				
Test results:				
Report distribution of disease severity				
Report cross-tabulation of index tests by results of reference standard				
Estimates:				
Report at least 4 estimates of diagnostic accuracy				
Diagnostic Accuracy Measures Reported:				
• Sensitivity				
• Spec				
• PPV				
• NPV				
• Likelihood ratios				
• Kappa				
• Area under the ROC curve / c-statistic				
• Accuracy/agreement				
• Other (specify)				
Report accuracy for subgroups (e.g. age, geography, different sex, etc.)				
If PPV/NPV reported, ratio of cases/controls of validation cohort approximate prevalence of condition in the population				
Report 95% confidence intervals for each diagnostic measure				
DISCUSSION:				
Discuss the applicability of the validation findings				

PRISMA-P (Preferred Reporting Items for Systematic review and Meta-Analysis Protocols) 2015 checklist: recommended items to address in a systematic review protocol*

Section and topic	Item No	Checklist item	Page
ADMINISTRATIVE INFORMATION			
Title:			
Identification	1a	Identify the report as a protocol of a systematic review	Page 1
Update	1b	If the protocol is for an update of a previous systematic review, identify as such	
Registration	2	If registered, provide the name of the registry (such as PROSPERO) and registration number	Page 2: Trial registration number PROSPERO 2015 CRD42015029204
Authors:			
Contact	3a	Provide name, institutional affiliation, e-mail address of all protocol authors; provide physical mailing address of corresponding author	Page 1
Contributions	3b	Describe contributions of protocol authors and identify the guarantor of the review	Page 10
Amendments	4	If the protocol represents an amendment of a previously completed or published protocol, identify as such and list changes; otherwise, state plan for documenting important protocol amendments	At this stage there are no relevant amendments to perform
Support:			
Sources	5a	Indicate sources of financial or other support for the review	Page 10 *not funded*
Sponsor	5b	Provide name for the review funder and/or sponsor	Page 10
Role of sponsor or funder	5c	Describe roles of funder(s), sponsor(s), and/or institution(s), if any, in developing the protocol	Page 10
INTRODUCTION			
Rationale	6	Describe the rationale for the review in the context of what is already known	Page 4 and 5
Objectives	7	Provide an explicit statement of the question(s) the review will address with reference to participants, interventions, comparators, and outcomes (PICO)	Page 5
METHODS			
Eligibility criteria	8	Specify the study characteristics (such as PICO, study design, setting,	Page 5, 6:

		time frame) and report characteristics (such as years considered, language, publication status) to be used as criteria for eligibility for the review	
Information sources	9	Describe all intended information sources (such as electronic databases, contact with study authors, trial registers or other grey literature sources) with planned dates of coverage	Page 6.
Search strategy	10	Present draft of search strategy to be used for at least one electronic database, including planned limits, such that it could be repeated	Appendix 1 in Supplemental file
Study records:			Page 7:
Data management	11a	Describe the mechanism(s) that will be used to manage records and data throughout the review	Page 7:.
Selection process	11b	State the process that will be used for selecting studies (such as two independent reviewers) through each phase of the review (that is, screening, eligibility and inclusion in meta-analysis)	Page 7:
Data collection process	11c	Describe planned method of extracting data from reports (such as piloting forms, done independently, in duplicate), any processes for obtaining and confirming data from investigators	Page 7:
Data items	12	List and define all variables for which data will be sought (such as PICO items, funding sources), any pre-planned data assumptions and simplifications	Page 7/8 .
Outcomes and prioritization	13	List and define all outcomes for which data will be sought, including prioritization of main and additional outcomes, with rationale	Page 5
Risk of bias in individual studies	14	Describe anticipated methods for assessing risk of bias of individual studies, including whether this will be done at the outcome or study level, or both; state how this information will be used in data synthesis	Not applicable. The present review will apply the STARD criteria (Appendix 2 in Supplemental file).
Data synthesis	15a	Describe criteria under which study data will be quantitatively synthesised	No cumulative evidence will be presented.
	15b	If data are appropriate for quantitative synthesis, describe planned summary measures, methods of handling data and methods of combining data from studies, including any planned exploration of consistency (such as I^2, Kendall's τ)	
	15c	Describe any proposed additional analyses (such as sensitivity or subgroup analyses, meta-regression)	
	15d	If quantitative synthesis is not appropriate, describe the type of summary planned	
Meta-bias(es)	16	Specify any planned assessment of meta-bias(es) (such as publication	Not applicable

		bias across studies, selective reporting within studies)	
Confidence in cumulative evidence	17	Describe how the strength of the body of evidence will be assessed (such as GRADE)	The present review will apply the STARD criteria. Page 8 . (Appendix 2 in Supplemental file).

*** It is strongly recommended that this checklist be read in conjunction with the PRISMA-P Explanation and Elaboration (cite when available) for important clarification on the items. Amendments to a review protocol should be tracked and dated. The copyright for PRISMA-P (including checklist) is held by the PRISMA-P Group and is distributed under a Creative Commons Attribution Licence 4.0.**

From: Shamseer L, Moher D, Clarke M, Ghersi D, Liberati A, Petticrew M, Shekelle P, Stewart L, PRISMA-P Group. Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015: elaboration and explanation. BMJ. 2015 Jan 2;349(jan02 1):g7647.